UNIVERSITA' DEGLI STUDI DI VERONA

DEPARTMENT OF

Biotechnology

GRADUATE SCHOOL OF

Natural Sciences and Engineering

DOCTORAL PROGRAM IN

Biotechnology

XXIX cycle

TITLE OF THE DOCTORAL THESIS

ANALYSIS AND INTERPRETATION OF WHOLE EXOME SEQUENCING DATA OF LEUKEMIA PATIENTS

S.S.D. BIO/18

Coordinator: Prof. Massimo Delledonne

Tutor: Prof. Massimo Delledonne

Doctoral Student: Dott./ssa Marianna Garonzi

ABSTRACT

Leukemias are a cancer type which affects the leukocytes progenitor cells. These malignancies are highly heterogeneous in terms of molecular mechanisms involved in their onset and progression. Heterogeneity can be further observed within the same subgroup of disease at the inter-individual level, being reflected by different clinical outcomes and responses to treatment in different patients. Unfortunately, the exact leukemia aetiology is still poorly understood and consequently also related prevention, diagnostic, prognostic and follow up methods remain mainly unidentified. Therefore, early-diagnosis, together with specifically tailored approaches to leukemia treatment, still represents a key point in determining patients' health, life quality and estimated life. Several efforts have been started to improve diagnosis, treatment and disease monitoring of leukemia. In this regard, the work presented in my PhD thesis is part of an international project, named "NGS-PTL: Next Generation Sequencing platform for targeted Personalized Therapy of Leukemia", whose objective is the development of technologies for the diagnosis and prognosis of haematological cancers. According to the project's objective, my thesis work aims to identify sequence variants from Whole Exome Sequencing data for the acute types of leukemia, to be used as potential biomarkers to improve therapeutic interventions and for personalize treatments. The work describes the setup and application of a bioinformatic pipeline able to identify the somatic mutations in the leukemia patients and the driver carrier genes, again with the result obtained by its application on all the samples of the project.

The setup of the pipeline has required the identification of a set of tools to apply to Cancer sequencing data. In particular, selection of dedicated software to perform the initial pre-processing of the data guarantees the use of sequencing data of high quality and ensures that the subsequent analysis will be performed on well-generated data. Moreover, the selection of MuTect as variant caller has

allowed us to overcome specific problems related to the heterogeneity of Cancer sample. The application of these software has led us to the identification of a large and reliable set of somatic variants to be evaluated for the identifications of new biomarkers and driver genes. Then, the interpretation of the somatic variants has required the use of specific database and resources to correctly interpret them and eventually to correlate the mutations with the driving or the development of the leukemia. Using the available biological knowledge, we were able to select likely highly damaging variants, some of which already connected with leukemia in cancer-related sources (COSMIC, ICGC and CIViC). At the end, the discover of genes that drives the development of the disease was performed using three statistical tools on the set of annotated mutations for each leukemia type, leading to the identification of a total of 32 biomarkers. In conclusion, the discovery of potential novel biomarkers, again with the additional biological information provided by the specific resources applied has demonstrated the importance of the application of NGS in the study of Leukemic patients.

CONTENTS

INTRODUCTION

LEUKEMIA

The term “leukemia” represent a group of cancers which affects the leukocytes progenitor cells. This malignancy occurs when alterations in the normal regulatory processes leading to blood cells development causing uncontrolled proliferation and differentiation arrest of hematopoietic stem cells in the bone marrow.

HEMATOPOIESIS

Blood cells formation, also called hematopoiesis, is driven by hematopoietic stem cells, and occurs in the bone marrow. Hematopoietic stem cells are pluripotent progenitor cells with the capacity of self-renewal and differentiation. The formation of mature and functional blood cells occurs via several consecutive cell divisions and maturation stages. In particular, hematopoietic stem cell can produce blood cells following two main different lineages, one represented by myeloid stem cells and the other by lymphoid stem cells (Figure 1):

- Myeloid cells: myeloid stem cells can generate red blood cells and platelets. In alternative, they differentiate to myeloblasts, immature cells of myeloid origin. Myeloblasts can produce several types of white blood cells known as granulocytes, a lineage that includes neutrophils, eosinophils, and basophils.
- Lymphoid cells: lymphoid stem cells differentiate to lymphoblasts, which can produce several types of white blood cells that are different from granulocytes: B lymphocytes, T lymphocytes and Natural killer cells.

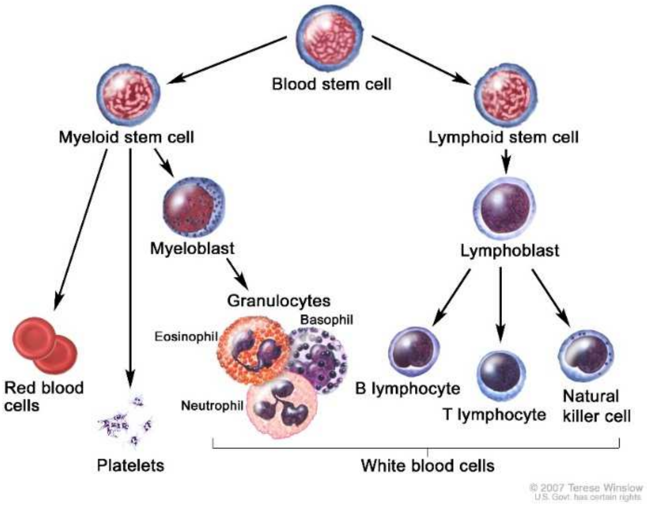


Figure 1. Blood cell development. Blood stem cells go through several intermediate steps to generate red blood cells, platelets, or white blood cells. Taken from www.cancer.gov

Blood circulates through the arteries and veins with all blood cell types, namely red blood cells, white blood cells and platelets, which perform different functions throughout the body. Red blood cells, also called erythrocytes, make up about 40 to 50 percent of the total blood volume. Red blood cells live for approximately 120 days before being replaced by new cells produced in the bone marrow. These cells contain a protein called haemoglobin, which carries oxygen throughout the body and deliver carbon dioxide from tissues to the lungs to be exhaled. Platelets, also called thrombocytes, are cell fragments rather than whole cells. They clump together to form clots in case a blood vessel wall is damaged. Clots traps also red blood cells and act as plugs to stop bleeding and serve as a base for healing of the injured area and tissue renovation. White blood cells, also called leukocytes, are much fewer in number than red blood cells. These cells constitute the human immune system. The 5 different subsets of white blood cells work together to protect the body by attacking foreign invaders, as bacteria and viruses, and endogenous dysregulated cells as tumors.

LEUKEMIA CELLS

In leukemia, the bone marrow produces abnormal white blood cells called leukemic cells. Leukemic cells are characterized by an altered differentiation status and a dysregulated cell cycle. As a consequence, the production of these cells alters the physiological composition and life-cycle of blood cells (formation, growth, function and death) thus impairing the ability of the bone marrow to produce normal blood cells. Moreover, because of their dysregulated cell cycle, leukemia cells do not die normally when they become old or damaged but accumulate abnormally and crowd out the healthy blood cells. Thus, over time, the continue increasing number of Leukemic cells alter the normal blood function including its oxygen capacity, the ability to control bleeding and fight infections.

TYPES OF LEUKEMIA

Leukemias are highly heterogeneous malignancies both in terms of phenotypes and molecular mechanisms underlying their onset and progression. Heterogeneity can be further observed within the same subgroup of disease at the inter-individual level, and reflects in different clinical outcomes and responses to treatment. There are several ways to categorize the leukemias based on different criteria. One of these is the classification of leukemias on the basis of the affected tissues (Figure 2):

- Myeloid Leukemia: originates from myeloid cells and it is called myeloid, myelogenous, or myeloblastic leukemia.
- Lymphoid Leukemia: originates from lymphoid cells and it is called lymphoid, lymphoblastic, or lymphocytic leukemia.

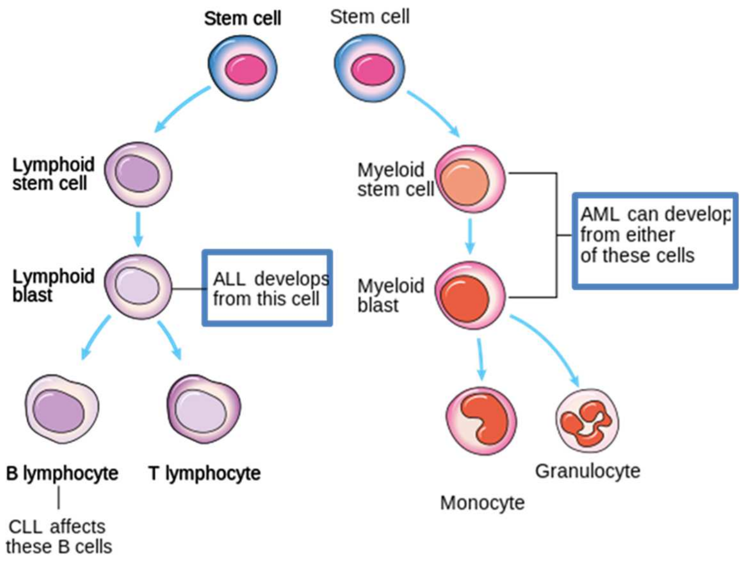


Figure 2. Leukemia types. Modified from www.cancerresearchuk.org

Leukemias can be further classified based on how quickly the disease develops and worsens:

- Acute: Acute leukemia is a fast-growing cancer that usually worsen quickly, if not treated. The abnormal blood cells composing the acute leukemia are very immature blasts (lymphoblasts) that grow rapidly and cannot carry out the normal functions of the white blood cells they derive from.
- Chronic: Chronic leukemia is a slower-growing cancer that worsen slowly over time. The number of abnormal blasts produced is low and, in general, these cells composing this type of leukemia are more mature and maintain some of the normal functions of myeloid cells.

According to these classifications, leukemias can be sub-grouped in four main types:

- Acute lymphocytic leukemia (ALL) is a condition where the bone marrow produces large numbers of abnormal immature lymphocytes (lymphoblasts). ALL can be further subdivided in different subsets. For example, on the basis of the lineage that the abnormal lymphoblasts originate from, as immature B or T lymphocytes (B-ALL or T-ALL, respectively). Typically, ALL develops quite quickly (acutely) and rapidly becomes worse (over a few weeks or so) unless treated.

- Acute myeloid leukemia (AML) is a condition where the bone marrow produces large numbers of abnormal immature white blood cells which are derived from a myeloid stem cell (myeloblasts). AML can be further subdivided on the basis of what cell type they derive from and their maturation stage. There are eight main subtypes of AML: M0, M1, M2, etc, up to M7. Typically, AML develops quite quickly (acutely) and rapidly becomes worse (over a few weeks or so) unless treated.
- Chronic lymphocytic leukemia (CLL) is a condition where a subject has an abnormal number of dysregulated B lymphocytes. The lymphocytes look phenotypically normal, e.g. features visible under a microscope, but they do not function properly. The main reason for the accumulation of abnormal lymphocytes is because they have a longer life-spam as compared to normal lymphocytes Typically, CLL progresses very slowly over months or years, even without any treatment.
- Chronic myeloid leukemia (CML) also known as chronic granulocytic leukemia (CGL) develops due to the accumulation of an abnormal stem cell subset of myeloid origin. As a consequence, there is also an expansion of the cells that originate from the abnormal myeloid progenitor, i.e. neutrophils, basophils and eosinophils, that develop into nearly-normal white cells, but over-accumulate in the bloodstream. Typically, CML develops and progresses slowly over months or years, even without treatment.

Despite a preliminary diagnosis of leukemia can be made with a simple complete blood count, extensive testing is required to differentiate myeloid and lymphoid leukemia and chronic versus acute leukemia. The treatment and prognosis of these malignancies are extremely different between the various types of leukemias. Moreover, as an early treatment provides the best opportunity for cure, the fast and accurate diagnosis of the right subtype of the disease is essential.

THE GENETICS OF LEUKEMIA

In the last decade, leukemia, as well as other cancers, have been proven to be essentially a condition of aberrant genetic programming [1], where changes of the genomic sequence in specific cells alter the structure, function, and/or expression of proteins that control their homeostatic processes, including cell growth, proliferation, differentiation, and apoptosis. The dysregulation of these critical functions ultimately leads to neoplastic transformation.

As general mechanism, cancer is the result of changes occurred in the DNA sequence of the genome of cancer cells [2]. Human cells normally acquire random mutations during the course of a person's life, and typically the human body is able to correct most of them. However, the continuous acquisition of genetic variations in individual cells may lead to the acquisition of deleterious mutations that confer the capability to proliferate and survive, causing the uncontrolled development of cancer.

The set of differences acquired in the DNA of a cancer cell genome are called somatic mutations, to distinguish them from germline variants which are inherited from parents and are transmitted to the progeny. Also, as not all the acquired abnormalities are effectively involved in the development of cancer, somatic mutations can be differentiated between two groups, named 'driver' and 'passenger' mutations (Figure 3). A *driver mutation* is a mutation directly implicated with the development of cancer by conferring growth advantage to the cancer cell, while *passenger mutations* do not confer clonal growth advantage and, therefore, do no contribute to cancer development. Cancer subsequently evolves through cycles of clonal expansion, that leads to further genetic diversification and clonal selection. As clones and subclones expand selective pressures can ultimately generate a highly variable patterns of genetic diversity [3]. This mechanism is also implicated in development of resistance to drugs through selection of resistant variants and is the primary cause of therapeutic failure.

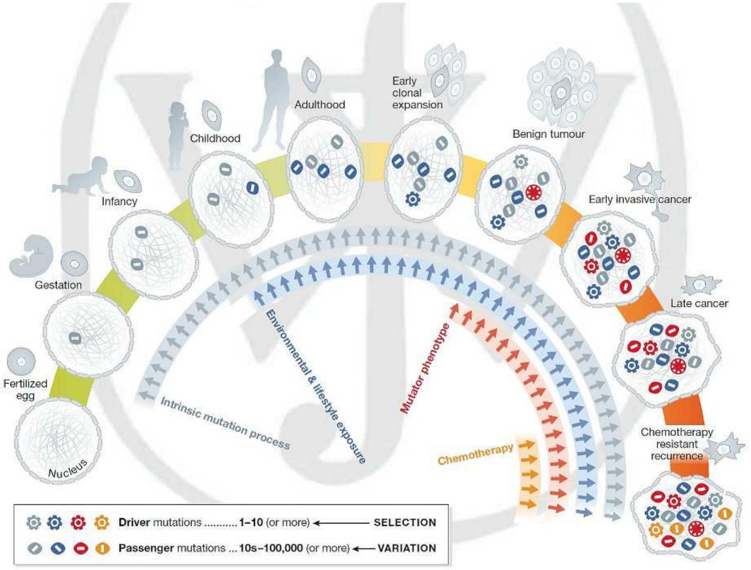

Fertilized egg
Nucleus
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late cancer
Chemotherapy resistant recurrence
Intrinsic mutation process
Environmental & lifestyle exposure
Mutator phenotype
Chemotherapy
Driver mutations ....1–10 (or more) ← SELECTION
Passenger mutations ...10s–100,000 (or more) ← VARIATION

Figure 3. The cellular lineage between a fertilized egg and a fully malignant cancer cell. [4]

The genetic aberrances that can be found in leukemic cells are highly diverse and varies between the different type of leukemia. These aberrances include chromosomal changes like the *translocation,* that are caused by chromosomes that swap some of their DNA, leading to a part of one chromosome becomes attached to part of a different chromosome. Other types of chromosome changes include the *inversion,* which means that a part of a chromosome is in reverse order, or a *deletion* that indicates a partial loss of a chromosome, or a *duplication* of a chromosome or a part of it. However, not only chromosome changes but also single nucleotide alteration concurs in determining the patient outcome and the development of the disease.

Among the genetic aberrances that can be found in leukemia, there are several that characterize the development of a specific type of leukemia. CML, for example, is characterized by the presence of the Philadelphia chromosome, a translocation between chromosomes 9 and 22 in humans, resulting in a fusion between the 5' end of the BCR gene and the 3' end of the ABL1 gene [5]. Although the Philadelphia chromosome may be found in other types of leukemias, presence of a BCR-ABL1 fusion gene is an absolute diagnostic criterion for CML. Another type of leukemia, the CLL, is instead characterised by a different set of genetic lesions that are typically the 13q deletions (55%; associated with favourable clinical outcome), trisomy 12 (15%; associated with intermediate prognosis), 11q deletions (12%; associated with poor clinical outcome), 17p deletions (8%; associated with poor clinical outcome), and recurrent mutations (2–11%) in NOTCH1, SF3B1, BIRC3, TP53, and MYD88 [6], [7].

The acute types of leukemia is more complex in terms of the genetic mechanisms of their development. AML can occur with somatic changes affecting some specific types of cells through a "two-hit" process. In other words, for leukemogenesis to occur, two types of mutations, or "two hits," are needed: 1) a mutation that improves hematopoietic cells' ability to proliferate (class I, including FLT3 and KIT), and 2) a mutation that prevents the cells from maturing (class II, including CBFB-MYH11, CEBPA, DEK-NUP214, MLL-MLLT3, NPM1, PML-RARA, RUNX1-RUNX1T1; [8], [9]). However, AML is the most clinically and biologically heterogeneous type of leukemia, and as study of genetic variation in AML continues, the aetiology of this disease is continuously being modified and integrated with new types of mutations, including mutations in epigenetic modifiers such as IDH1, IDH2, and DNMT3A. Moreover, also ALL is characterized by complex types of structural rearrangements, copy number alterations, and mutations in specific genes (i.e. gene regulating lymphoid development). Approximately 20% of B-ALL cases harbour genetic alterations that activate kinase signalling, including rearrangements of the cytokine receptor gene CRLF2; rearrangements of ABL1, JAK2, and PDGFRB; and mutations of JAK1 and JAK2. Other class of mutation includes hematopoietic regulators (ETV6 and RUNX1), tyrosine kinases, and epigenetic regulators [10]. Both in AML and ALL there is a lot of knowledge still to uncover under the genetic variability of these condition.

Since Nowell and Hungerford identified the t(9;22) translocation (the Philadelphia chromosome) associated with chronic myeloid leukemia, a wealth of data has accumulated showing that the karyotype and mutation status of certain genes provide important prognostic, and in some cases, therapeutic information for leukemia. There are several prognostic factors that are determined by cytogenetics; more specifically, by acquired mutations that, once detected, make it possible to define the appropriate treatment for a given patient.

Specific aberrations are used for patient risk stratification and to guide the patient management, ad correlate with favourable and unfavourable outcome (Table 1).

Response Rate	French American British classification	Karyotype	Molecular Change
Low	M4, M5	t(6;11)(q27;q23)	AF6(6q27)
Low	M4, M5	t(10;11)(p12;p23)	AF10(p12)
Low	M5	t(11;17)(q23;q21)	ALL 1(11q23)
Low	M4, M5	t(11:19)(q23;p13)	ELL(19p13.1)
Low	M1, M2, M4, M6	t(3;3)(q23:q26)	Gene activation
Low	M0, M1, M4, M5, M6, M7	inv(3)(q21;q26)	Gene activation
Low		5;5q-	
Low		7;7q-	
Low	L1	t(1:19)(q23;p13)	E2A, PBX1
Low	L3	t(8;14)(q24;q11)	
Moderate	L3	t(8;14)(q24;Q32)	IGH, cMYC
High	M2	t(8;21)(q22;q22)	ETO (8q22)
High	L1	t(9;22)(q34;q11)	cABL,BCR
High	L1	t(4;11)(q21;q23)	MLL, AF4

Table 1. Leukemia karyotypes and molecular changes associated with response rate

Despite increasing knowledge of the effects of genetic variation on prognosis of leukemia, these are only just few examples of genomic alterations that are related to the leukemia outcome. Many others have already been detected but the majority of mutations that drive the development of leukemias are still not known, and there are few options for tailoring treatment based on known genetic characteristics. Therefore, mutation discovery using genome-wide strategies recently became the state-of-art approach to investigate the genetic alterations linked to leukemia, as it provides a non-biased way to identify novel causative mutations underlying leukocyte dysregulation. Challenges for the future are to comprehensively identify and experimentally validate all genetic alterations driving leukemogenesis and treatment failure in leukemia and to implement genomic profiling into the clinical setting to guide risk stratification and targeted therapy.

NEXT GENERATION SEQUENCING APPLIED TO LEUKEMIA DIAGNOSTICS

Next-generation sequencing (NGS) provides the basis for the identification of novel diagnostic and therapeutic strategies as it makes the sequencing of individual genomes accessible at a reasonable cost. During the last decade, due to the continuous development of sequencing technologies, the cost for sequencing a human genome has decreased to only about 1000$. This means that the sequencing technology can be used for the discovery of medically relevant variations present in individual patients as well as the fast and cost-efficient assessment of the genetic variability within cohorts of patients affected by the same disease.

NGS technology provides an unprecedented view of genome sequence and alterations down to the single-base resolution. NGS is also extremely flexible as it allows to investigate either the complete genomic sequence in whole-genome sequencing (WGS) or to focus on specific genomic regions of interest, such as protein coding genes in whole-exome sequencing (WES). In particular, WES has been widely used in clinical studies as it allows to concentrate on highly informative exonic sequences. Even if the exome represents less than 2% of the human genome, it is the most crucial component as mutations in the exome can directly affect the protein structure and function and most likely result in clinical phenotypes. Not surprisingly the exome contains about the 85% of known disease-causing variants [11]. Moreover, WES is far cheaper than the WGS, allows a higher number of samples to be analysed per sequencing run and is thus more suitable to the analysis of larger cohorts of clinical samples.

To sequence only the exons of a genome the DNA has to be processed following some basic steps, as shown in Figure 4:

1. The genomic DNA is randomly sheared to construct an *in vitro* shotgun library. The library fragments are also ligated to adaptors to allow the subsequent sequencing.
2. The library is enriched for sequences corresponding to exons (dark blue fragments) by aqueous-phase hybridization capture: the fragments are hybridized to biotinylated DNA or RNA baits (orange fragments) in the presence of blocking oligonucleotides that are complementary to the adaptors.
3. Recovery of the hybridized fragments by using streptavidin-conjugated beads that can bind the biotins presents on the probes. The capture fragments are then amplified and sequenced in an NGS instrument.
4. Reads are mapped on a reference genome and candidate somatic variants are identified.

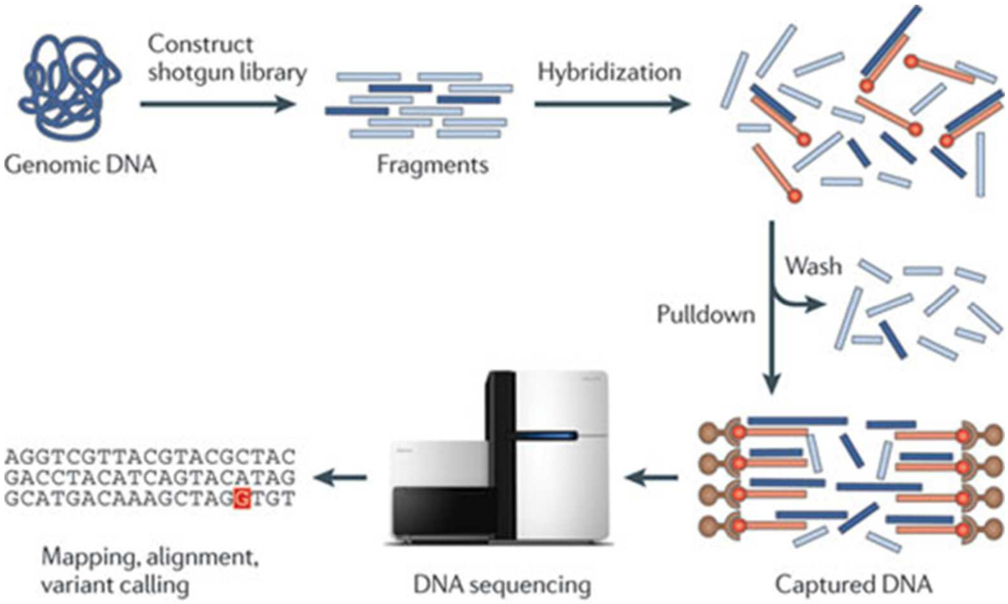


Figure 4. Workflow for exome sequencing

IDENTIFICATION OF SOMATIC VARIANTS

The process that goes from sequencing data to a reliable set of somatic mutations is complicated by the presence of confounding factors such as sequencing errors, misalignments or repetitive sequences. To ensure the accurate detection of somatic variants it is necessary to perform several pre-processing of the sequenced reads. The step of pre-processing includes the removal of reads derived from PCR duplicates, the filtering of low quality reads and the removal of adaptor sequences. Then, methods specifically dedicated to the identification of somatic mutations must be applied. These methods should implement stringent filtering to remove false positives due to high GC content, strand bias (reads indicating a possible mutation only align to one DNA strand) or from poor mapping resulting from repetitive or low complexity sequence in the reference genome.

Most tumor samples, including leukemic cells, are a heterogeneous collection of cells, containing both normal and cancerous cells thus further challenging the identification of somatic mutations. Therefore, dedicated analysis methods should be applied to detect low frequency variants that represent the cancer cells, within the high background signal due to cells with normal genome. Standard variant callers are based on the assumption of a diploid genome in which variants are either present in heterozygous or homozygous state. This model does not apply when only a limited portion of cells in the sample show the variant. As a result, most of real somatic variants are just discarded as background noise. Different approaches have been thus implemented [12]–[14]. Among these, MuTect software has been successfully used to identify somatic mutations in mixed samples and is a widely-recognized method for somatic variant calling in cancer samples. While the majority of existing methods typically miss low-allelic-fraction mutations that occur in only a subset of the sequenced cells owing to either tumor heterogeneity or contamination by normal cells, MuTect is specifically created to detect subpopulations of variants with very low allele fractions (10%) and only a few reads supporting somatic mutations.

MuTect takes as input the sequence data from the tumor and the matched normal DNA after alignment of data to a reference genome and standard pre-processing steps. MuTect applies a statistical analysis that identifies high confidence sites that are likely to carry somatic mutations. The analysis predicts a somatic mutation by using two Bayesian classifiers: the first aims to detect whether the tumor is non-reference at a given site; for those sites that are found as non-reference, the second classifier makes sure that the normal sample does not carry the variant allele. In practice the classification is performed by calculating a LOD score (log odds) and comparing it to a cutoff determined by the log ratio of prior probabilities of the considered events.

For the tumor:

$$LOD_r = \ log_{10}\left(\frac{P(observed\ data\ in\ tumor\ |site\ is\ mutated)}{P(observed\ data\ in\ tumor\ |site\ is\ reference)}\right)$$

For the normal:

$$LOD_N = \ log_{10}\left(\frac{P(observed\ data\ in\ normal\ |site\ is\ reference)}{P(observed\ data\ in\ normal\ |site\ is\ mutated)}\right)$$

Since the somatic mutations are expected to occur at a rate of ~1 per Mb, are required that $LOD_r > log_{10}(0.5x10^{-6})≈6.3$ which guarantees that the false positive rate, due to noise in the tumor, is less than half of the somatic mutation rate. In the normal tissue, since germline variants occur roughly at a rate of 100 per Mb, are required that $LOD_N > log_{10}(0.5x10^{-2})≈2.3$. This cutoff guarantees that the false positive rate of the somatic call, namely due to the missing identification of the variant in the normal, is also less than half the somatic mutation rate.

ANNOTATION AND PRIORITIZATION OF SOMATIC VARIANTS

The first important step to assess the biological impact of a somatic mutation is to annotate it with the existing knowledge. In the context of exome sequencing, the annotation procedure starts with the identification of the protein-coding genes in which the variant is located and the assessment of their impact on the final protein product (Figure 5, Table 2).

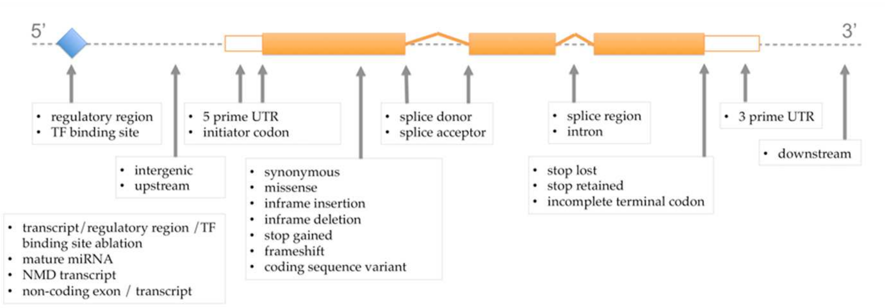


Figure 5. A diagram showing the location of each type of variant

Loss of Function	The variant is likely to cause the transcript's product to lose function. The ontologies included in this category are: transcript ablation, exon loss variant, stop lost, stop gained, initiator codon variant, frameshift variant, splice acceptor variant, splice donor variant
Missense	The variant will cause at least one amino acid to change or cause a premature start codon in the UTR5. The ontologies included in this category are: disruptive inframe deletion, disruptive inframe insertion, inframe deletion, inframe insertion, 5 prime UTR premature start codon gain variant, missense variant
Other	The variant is likely to have a low or unknown effect on the transcript's functional product. These changes do not change the amino acid sequence of the protein. The ontologies included in this category are: synonymous variant, stop retained variant, splice region variant, 3 prime UTR variant, 5 prime UTR variant, intron variant, non-coding exon variant, intergenic variant

Table 2. The categories of effect among the variant transcript interaction and the likely effect that the variant will have on the protein's product, including the ontologies that correspond to each effect category

In particular, mutations that can affect the function of a protein are the non-synonymous mutations. These include for example stop gain and frameshift mutations that by truncating the protein product may result in the inactivation of the protein. Also, missense mutations, which cause an aminoacidic sequence alteration, may also have an effect on protein function by altering its 3D structure or affecting its active site or regulatory sites. The assessment of the impact and the potential pathogenicity of these non-synonymous variants is the most crucial step in the annotation procedure and relies in the application of several computational methods. Prediction tools, such as SIFT [15], PolyPhen [16], MutationTaster [17], MutationAssessor [18] and GERP [19], have been developed to estimate whether a given variant is likely to be deleterious for the function of the encoded protein and are based on different principles, like the conservation among species, the biochemical properties of the encoded amino acids and the three-dimensional calculations of the protein structure. Moreover, one of the most effective ways to enrich a somatic variants dataset for the most-likely damaging variants is to use a population frequency filter, based on the concept that causative variants are rare and therefore not common within a reference healthy population. Several databases such as ExAC (Exome Aggregation Consortium) [20], the 1000 Genome Project [21] and the NHLBI Exome Sequencing Project (ESP6500) [22] provide population-level variant frequencies thus allowing to discriminate between innocuous common variants and potentially dangerous rare variants.

A further step for annotating and prioritizing variants is to use knowledge coming from previous studies. Several dedicated resources like the *Catalogue Of Somatic Mutations In Cancer* (COSMIC) [23], [24], the *International Cancer Genome Consortium* (ICGC) [25] data and the *Clinical Interpretation of Variants in Cancer* (CIViC) [26] database provide information about the recurrence of somatic mutations in cancer types and about known susceptibility/resistance to drugs associated to particular mutations.

IDENTIFICATION OF DRIVER GENES

Only a small subset of the somatic mutations found in cancer cells are responsible for tumorigenesis. The distinction of real driver mutations from passenger mutations is the most important task in cancer genome sequencing projects, and implies the identification of genes that exhibits signals of positive selection across a cohort of tumor samples. Among all the different approaches utilized at this aim, the most intuitive consists in the identification of genes that are mutated more frequently than expected given a certain background mutation rate. A second approach is based on the observation that driver mutations tend to clusterize in particular regions of the proteins, like for example kinase domains. Also, this second method exploits positive selection signals over the background mutation rate to identify genes containing putative driver mutations. While these methods are useful to identify highly recurrent driver genes and mutations, both are intrinsically limited in detecting lowly recurrent drivers. A third complementary approach has been developed which evaluates the functional impact of the mutations on the protein. This method detects putative driver genes by identifying those mutations biased towards higher functional impact. This approach doesn't rely on the estimation of a background mutation rate and is thus not limited to highly recurrent mutations. However, being based on assessment of the functional impact of mutations, it is generally more suited to the identification of loss of function events.

Clearly, no method can provide a comprehensive identification of driver genes due to intrinsic limitations. Thus, the combination of several approaches should be exploited to obtain the most comprehensive list of driver genes.

The next paragraphs are dedicated to the description of the three software selected for identification of driver genes in the present study which are based on the principles outlined above. They all require somatic variants data generated from a cohort of tumor samples.

MutSigCV

The first software selected, MutSigCV [27], works based on a recurrence-based approach to identify genes that are mutated more often than one would expect by chance. The method is based on the mutation frequency of an individual gene compared with the background mutation rate. The software corrects for possible variations by employing patient-specific mutation frequencies and mutation spectra (e.g., the percentages of mutations that are transitions of certain types, transversions of certain types, and/or nonsense), and gene-specific mutation rates, incorporating expression levels and replication times. Incorporating these covariates into the model substantially reduces the number of false positives in the generated list of significant genes.

The following figure (Figure 6) shows how the software works: on the left a set of chromosomes, each from the tumor of a different cancer patient. Genes are cartooned as coloured bands, and somatic mutations are indicated by red triangles. The mutations from all the tumors are aggregated together by merging the data from the different tumors, and the total number of mutations per gene can be computed. Then such tally is converted to a score, and then to a significance level. A threshold is chosen to control for the False Discovery Rate (FDR), and genes exceeding this threshold are reported as significantly mutated.

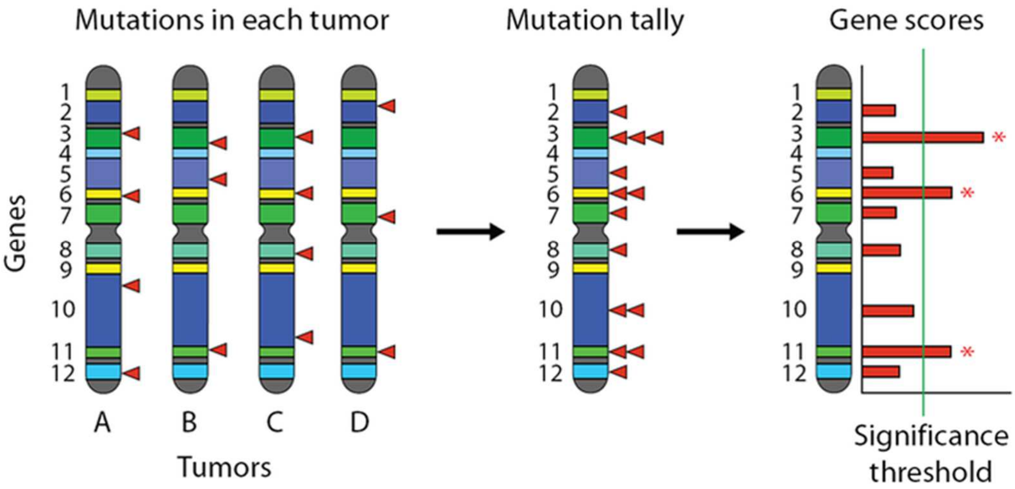


Figure 6. MutSigCV procedure

MutSigCV produces a report of significant genes, listed in descending order from the most significant to least significant ones.

ONCODRIVECLUST

The second software selected, OncodriveCLUST [28], has an approach based on mutation clustering on protein domains. The method is designed to exploit the fact that mutations in cancer genes, especially oncogenes, often cluster in particular positions of the protein and therefore do not occur with equal probability on all the positions of a gene (Figure 7). Clustering within specific regions suggests they mutations are positively selected during the clonal tumor evolution, and might therefore alter the function of the protein conferring an adaptive advantage to the cancer cells. Such feature can thus be exploited to nominate novel candidate driver genes.

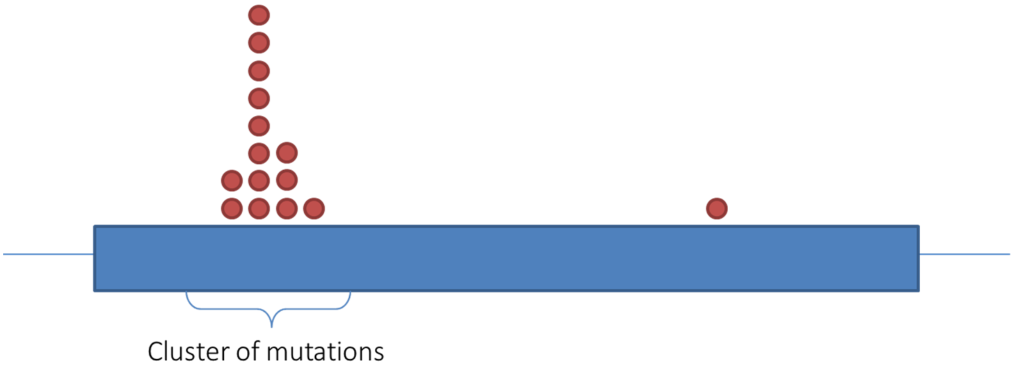


Figure 7. Mutation clustering on specific position of a gene

The method does not assume that the baseline mutation probability is homogeneous across all gene positions but creates a background model using silent mutations. Coding silent mutations are supposed to be under no positive selection and may reflect the baseline clustering of somatic mutations.

The software works by performing four main steps:

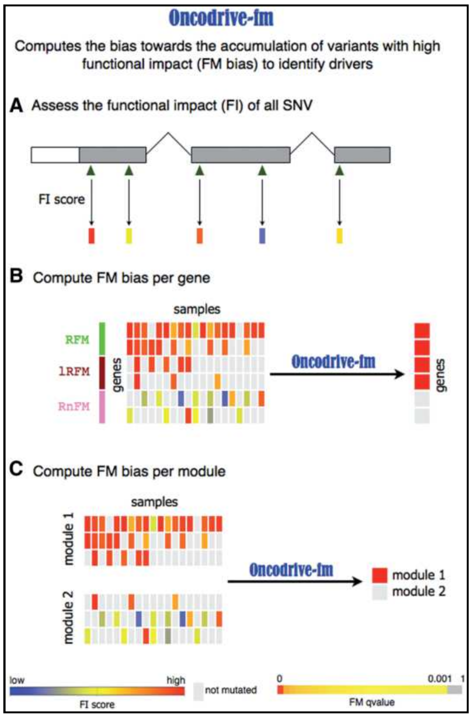
- mutations affecting proteins are clustered by gene across a cohort of tumor samples. Those protein residues having a number of mutations barely expected by chance are selected as candidate positions.
- these positions are grouped to form mutation clusters

- each cluster is scored with a figure proportional to the percentage of the mutations that are enclosed within that cluster, and inversely related to its length. The gene clustering score is obtained as the sum of the scores of all clusters (if any) found in that gene
- each gene clustering score is compared with the background model to obtain a significance value. The background model is obtained performing the same steps than above but assessing only coding silent mutations.

OncodriveFM

The last software selected, OncodriveFM [29], is based on the identification of the functional impact of variants. It computes a metric of functional impact using three well-known methods (SIFT, PolyPhen2 and MutationAssessor) and assesses how much the functional impact of variants found in a gene across several tumor samples deviates from a null distribution. OncodriveFM is thus based on the assumption that any bias towards the accumulation of variants with high functional impact is an indication of positive selection and can thus be used to detect candidate driver genes or gene modules and to prioritize genes or pathways.

The software starts by computing three metrics of functional impact (FI score) for each non-synonymous single nucleotide variants (nsSNVs) found in genes across a list of tumor samples (Figure 8). Stop-gain SNVs (stSNVs) and frameshift-causing indels (fsindels) are incorporated into the bias analysis by assigning them scores that are comparable to the highest-ranking tier of nsSNVs. Finally, synonymous SNVs (sSNVs) are taken into account with scores equal to those of bottom ranking nsSNVs. The second step starts by averaging the FI scores of variants per gene and comparing them to the distribution of scores of variants in functionally similar genes. if the somatic SNVs are obtained using a whole-exome sequencing approach, the null distribution contains the entire set of SNVs and fsindels detected across all tumor samples. The mean FI of each gene across all tumor samples is then probed for significance employing a permutation strategy.



Oncodrive-fm
Computes the bias towards the accumulation of variants with high functional impact (FM bias) to identify drivers
A Assess the functional impact (FI) of all SNV
FI score
B Compute FM bias per gene
samples
RFM
1RFM
RnFM
genes
Oncodrive-fm
genes
C Compute FM bias per module
samples
module 1
module 2
Oncodrive-fm
module 1
module 2
low
high
FI score
not mutated
0
0.001
1
FM qvalue

Figure 8. OncoDriveFM procedure

BACKGROUND OF THE PROJECT

Leukemia accounts for approximately 10% of the new diagnosed cancers every year, with an overall incidence that is slightly higher in subjects of European ancestry. Unfortunately, despite the huge advances in the clinical treatment of some subtypes of leukemia, many still have a poor prognosis. In addition, in a subset of long-term surviving patients, treatment results are unsatisfactory for short and long-term toxicities. Reason of this picture is that the exact leukemia aetiology is still poorly understood and consequently also related prevention, diagnostic, prognostic and follow up methods remain mainly unidentified. The early-diagnosis, together with specifically tailored approaches to leukemia treatment, still represents key points in determining patients' health, life quality and estimated life.

Several initiatives [30], thanks to collaborative groups and international projects, have been started to improve diagnosis, treatment and disease monitoring for leukemia. At this regard, my PhD project is part of a bigger international project, named NGS-PTL, "Next Generation Sequencing platform for targeted Personalized Therapy of Leukemia", financed by the European Union through the seventh framework program. The project involved 10 international partners in a multidisciplinary approach, comprising the fields of clinical medicine, industry research, NGS technology, molecular biology, genomics, transcriptomics, biostatistics and bioinformatics. The objective of NGS-PTL project was the development and validation of methods for the diagnosis and prognosis of haematological cancers. These included quality control and analytical tools, based on the most innovative massive parallel DNA/RNA sequencing technologies. The NGS-PTL project aimed to provide the basis for a completely new knowledge of leukemia aetiology and of the molecular mechanisms underlying inter-individual variability in response to treatments.

Uncovering the genomic variability among and within leukemia subtypes is of utmost importance to guide the therapeutic interventions on these diseases and constitutes the basis of the NGS-PTL project and of these work. In particular, the analysis reported here was focused on the main type of leukemia patients present in the project, that is the acute subtype of leukemia, the more complex in terms of the genetic mechanisms involved in their development.

AIM OF THE STUDY

In agreement with the NGS-PTL project’s objectives, my work aimed to identify sequence variants from Whole Exome Sequencing data of two different types of leukemia (AML and ALL), to select potential biomarkers of the disease to be investigated in future studies to improve therapeutic interventions and to tailor personalize treatments.

To obtain this result, the work performed during my PhD focused on the setup, validation and implementation of a bioinformatic pipeline to identify somatic mutations from WES data of leukemia patients and to select candidate driver carrier genes in the analysed samples.

MATERIALS AND METHODS

SELECTED SAMPLES

This work involved the analysis and interpretation of WES data derived from leukemia patients. The cohorts of patients selected belongs to two main types of Leukemia, the Acute Myeloid Leukemia (AML) and the Acute Lymphoblastic Leukemia (ALL). The selected patients and samples are summarized in Table 3.

Leukemia type	# samples	# patients
AML	128	64
ALL	77	37

Table 3. Number of samples and patients for each leukemia type selected in the project

To identify somatic variants characterizing the leukemia and unambiguously discriminate them from inherited germline variants, multiple samples corresponding to the control "normal" tissue (usually saliva) and the tumoral tissue (peripheral blood or bone marrow), collected at one or multiple timepoints (onset and relapse of disease), were sequenced for each patient. In particular, two cohorts of patients were selected for AML cases. The first cohort comprised 42 cases which included 4 patients with a normal karyotype, 25 patients with one or two chromosomal abnormalities and 13 patients with a complex karyotype, i.e. with more than two chromosomal abnormalities. 34 tumoral samples were collected at diagnosis and 8 samples at relapse along with their matched healthy control samples. In the second cohort 22 cases were selected, which include 6 cases with a complex karyotype and the remaining with one or two chromosomal abnormalities. All the samples were collected at the diagnosis and after the complete remission of the disease. For ALL, patients negative for the typical Philadelphia chromosome (BCR-ABL) translocation, as well as for other known recurrent molecular rearrangements (i.e. E2A-PBX, TEL, AML1-MLL-AF4), were selected. The matched tumoral/normal samples were collected from adult B-ALL patients at the time of diagnosis in 33 cases, at relapse in one case, at both diagnosis and relapse in 3 cases.

The preparation of the Whole Exome libraries was performed on all the 205 samples included in the study with two Illumina kits: the *TruSeq Exome Enrichment Kit* and the *Nextera Rapid Capture Exome* that are based on almost identical capture designs for the selection of exome sequences. Sequencing was performed using an Illumina HiSeq1000, generating sequencing reads of 100 nucleotides in paired end, i.e. every DNA fragment is sequenced twice, on the forward and reverse strand. Each genome region analysed was sequenced on average 80 time, i.e. 80X coverage, to ensure the detection of mutations associated with the disease at high sensitivity.

All the sequenced samples were analysed with the same workflow which can be divided into four parts:

1. preprocessing of raw reads obtained from WES and alignment to the reference genome sequence;
2. somatic variants calling;
3. variants annotation;
4. identification of driver-mutations carrier genes.

RAW READS PREPROCESSING AND ALIGNMENT TO REFERENCE GENOME

The preprocessing pipeline was based on a set of open source tools including different modules dedicated to data filtering, quality control (QC) and reads alignment, and is based on a well-established workflow [31] as summarized in Figure 9.

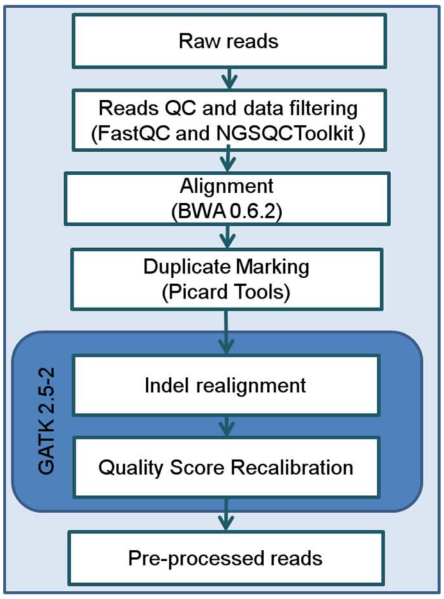

Raw reads
Reads QC and data filtering (FastQC and NGSQCToolkit)
Alignment (BWA 0.6.2)
Duplicate Marking (Picard Tools)
GATK 2.5-2
Indel realignment
Quality Score Recalibration
Pre-processed reads

Figure 9. The preprocessing pipeline

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and NGSQCToolkit [32] applications were selected to perform sequencing data QC and filtering. The FastQC software v. 0.10.1 was chosen to determine sequencing data quality before proceeding with the analyses as it provides a fast overview of the level of error of produced reads potentially affecting subsequent alignment and SNP calling steps. Then, it was chosen to add a filtering step to remove low quality reads and contaminant adaptor sequences, thus increasing the accuracy of results obtainable from produced data. For this purpose, the NGSQC toolkit was employed.

```
#FastQC (0.10.1)

fastqc --nogroup -t 2 sequence_1.fastq.gz sequence_2.fastq.gz -o FastQC

#ngsqctoolkit (2.3)

perl NGSQCToolkit_v2.3/QC/IlluQC_PRLL.pl -c 24 -t 2 -s 20 -l 70 -pe sequence_1.fastq.gz
sequence_2.fastq.gz 2 A -o sample_name/
```

For the alignment of the high quality paired-end reads to the hg19 reference genome the Burrows-Wheeler Aligner (BWA 0.6.2) was selected, a fast and memory-efficient read aligner widely used for WES alignment [33] that allows gapped alignment, thus enabling a more accurate alignment and detection also in correspondence of insertions and deletions (INDELs) [34]. The alignment data filtering was based on the Picard Tools (https://broadinstitute.github.io/picard/) to remove artifacts due to PCR duplicates.

```
#BWA (0.6.2)

bwa-0.6.2/bwa aln -t 24 ucsc.hg19.fasta sequence_1_filtered.fastq.gz >sequence_1_filtered.sai
bwa-0.6.2/bwa aln -t 24 ucsc.hg19.fasta sequence_2_filtered.fastq.gz >sequence_2_filtered.sai
bwa-0.6.2/bwa sampe -r
@RG\\tID:2\\tLB:flowcell\\tPL:illumina\\tSM:sample_name\\tPU:unk_barconde
ucsc.hg19.fasta sample_name/sequence_1_filtered.sai sample_name/sequence_2_filtered.sai
sample_name/sequence_1_filtered.fastq.gz sample_name/sequence_2_filtered.fastq.gz |
samtools view -Sbh - >sample_name/alignment.bam

#Picard (1.81)

java -Xmx16g -jar picard-tools-1.81/SortSam.jar VALIDATION_STRINGENCY=SILENT
TMP_DIR=sample_name/TMP MAX_RECORDS_IN_RAM=1000000
INPUT=sample_name/alignment.bam OUTPUT=sample_name/alignment_sorted.bam
SORT_ORDER=coordinate CREATE_INDEX=true
java -Xmx16g -jar picard-tools-1.81/MarkDuplicates.jar VALIDATION_STRINGENCY=SILENT
TMP_DIR=sample_name/TMP CREATE_INDEX=true REMOVE_DUPLICATES=true
ASSUME_SORTED=true INPUT=sample_name/alignment_sorted.bam
OUTPUT=sample_name/alignment_sorted_dedup.bam
METRICS_FILE=sample_name/alignment_sorted_dedud_duplicates.txt
```

Then, Genome Analysis Toolkit suite (GATK ver. 2.5.2) [35] was selected to perform local re-alignment and quality score recalibration. In more details: GATK was used to perform a local realignment of reads in correspondence of insertions and deletions to avoid false calls due to wrong alignments in "challenging" genomic regions. In particular, we realigned reads around known INDELs annotated in the 1000Genomes project dataset. GATK was also used to perform a recalibration of the quality of bases according to direct comparison with the reference genome, allowing to obtain more accurate results than simply relying on the base call accuracy measure provided by the sequencer. To avoid biases in the correction process, genomic positions corresponding to known variants annotated in dbSNP build 135 [36] were removed from the recalculation of base accuracy.

```
#GATK (2.5-2)

java -Xmx16g -jar GenomeAnalysisTK-2.5.2.jar -T IndelRealigner -R ucsc.hg19.fasta -I
alignment_sorted_dedup.bam -targetIntervals hg19.intervals -o output_realigned.bam -known
1000G_phase1.indels.hg19.orderchange.vcf -known dbsnp_135.hg19.orderchange.vcf --
consensusDeterminationModel KNOWNS_ONLY -LOD 0.4
java -Xmx16g -jar GenomeAnalysisTK-2.5.2.jar -T BaseRecalibrator -R ucsc.hg19.fasta -I
output_realigned.bam -o recalibrated.report -knownSites dbsnp_135.hg19.orderchange.vcf -
cov ReadGroupCovariate -cov QualityScoreCovariate -cov CycleCovariate
java -Xmx16g -jar GenomeAnalysisTK-2.5.2.jar -T PrintReads -R ucsc.hg19.fasta -I
output_realigned.bam -BQSR recalibrated.report -o recalibrated.bam

#NGSrich (0.7.8)

java -Xmx16g -cp NGSrich_0.7.8/bin/ NGSrich evaluate -r alignment_sorted_dedup.bam -u hg19
-a refGene.txt -t capture.bed -T TMP -o CAPTURE -p 2 -h 200 --no-details

#samtools (0.1.18)

samtools mpileup -d 100000 -q 0 -Q 0 -f ucsc.hg19.fasta alignment_sorted_dedup.bam -A
>alignment.mpileup
```

VARIANT CALLING

The variant calling pipeline was based on MuTect [14], a tool specifically created for the calling of somatic mutations in cancer samples. MuTect uses both dbSNP [36] and COSMIC [23], [24] to confidently call somatic variants by blacklisting common polymorphism in the population and retaining known mutations identified also in other cancer cases.

```
#VARIANT_CALLING_MuTect

java -Xmx2g -jar muTect-1.1.4.jar --analysis_type MuTect --reference_sequence
ucsc.hg19.fasta --cosmic b37_cosmic_v54_120711.chr.reorder.vcf --dbsnp
dbsnp_132_b37.leftAligned.chr.reorder.vcf --intervals all.intervals --input_file:normal
ctrl.bam --input_file:tumor tumor.bam --out call_stats.out --coverage_file coverage.wig.txt --
vcf variants.vcf
```

The tools, applied on all the patients' normal and tumoral samples, produces lists of candidate somatic mutations as variant calling format (vcf) files.

VARIANTS ANNOTATION

The annotation of putative somatic mutations was based on the VarSeq (http://goldenhelix.com/products/VarSeq/) software, a tool that provide variant discovery and interpretation for Next Generation Sequencing data, starting from vcf files.

VarSeq software were used to annotate and filter through the large variant data sets produced in the two different cohort of leukemia patients.

The annotation was based on the following databases:

- RefSeq Genes 105v2, NCBI [37]: defines genomic sequences to be used as reference standards for well-characterized genes. These sequences, labeled with the keyword RefSeqGene in NCBI's nucleotide database, serve as a stable foundation for reporting mutations, for establishing conventions for numbering exons and introns, and for defining the coordinates of other variations. Sequences of the RefSeqGene project provide stable gene-specific genomic sequence for each gene, as well as including upstream and downstream flanking regions.

- dbSNP132 [36]: The Database of Short Genetic Variations (dbSNP) is a repository of all types of short genetic variations less than 50 bp in length. dbSNP accepts submissions of common as well as polymorphic variations, and contains both germline and somatic variations. In addition to archiving molecular details for each submission and calculating submitted variant locations on each genome assembly, dbSNP maintains information about population-specific allele frequencies and genotypes, reports the validation state of each variant and indicates if a variation call may be suspect because of paralogy.
- 1000 Genomes - 1kG Phase3 [21]: this database contains variant frequencies from 1000 Genomes Project, and in particular minor allele frequency (MAF) for each subpopulation: Europeans, Asians, Africans and Admixed Americans, as well as a MAF field over all samples. These frequencies were calculated using 2,504 samples from the 1000 Genomes Project.
- NHLBI ESP6500 Exomes Variant Frequencies [22]: this databases contains variant frequencies from the NHLBI Exome Sequencing Project for each subpopulation: European Americans and African Americans, as well as a MAF field over all samples. These frequencies were calculates using 6503 samples from multiple ESP cohorts.
- ExAC Variant Frequencies 0.3, BROAD [20]: this database contains variant frequencies across a combined data set of 60,706 exomes of unrelated individuals belonging to 7 populations (i.e. NFE – Non-Finnish European) sequenced as part of various disease-specific and population genetic studies.
- CIViC - Variant Clinical Evidence [26]: a resource for Clinical Interpretation of Variants in Cancer. The database is a focused precision medicine resource for variants with published clinical evidence for the relationship

between given mutations and diagnosis, prognosis or response to a specific treatment of cancer.

- COSMIC [23], [24]: the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers.
- ICGC Simple Somatic Mutations [25]: a comprehensive catalogue of genomic abnormalities in tumors from different cancer types and/or subtypes which are of clinical importance.
- dbNSFP [38]: a database developed for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome. It compiles prediction scores from 18 prediction algorithms (SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, MutationTaster2, MutationAssessor, FATHMM, MetaSVM, MetaLR, CADD, VEST3, PROVEAN, FATHMM-MKL coding, fitCons, DANN, GenoCanyon, Eigen coding, Eigen-PC, M-CAP), 6 conservation scores (PhyloP x 2, phastCons x 2, GERP++ and SiPhy) and other related information including allele frequencies observed in the 1000 Genomes Project phase 3 data, UK10K cohorts data, ExAC consortium data and the NHLBI Exome Sequencing Project ESP6500 data, various gene IDs from different databases, functional descriptions of genes, gene expression and gene interaction information, etc.

Using a chain of filters based on the selected annotation sources is possible to narrow the list of variants down to those that are most likely to be of interest (Figure 10).

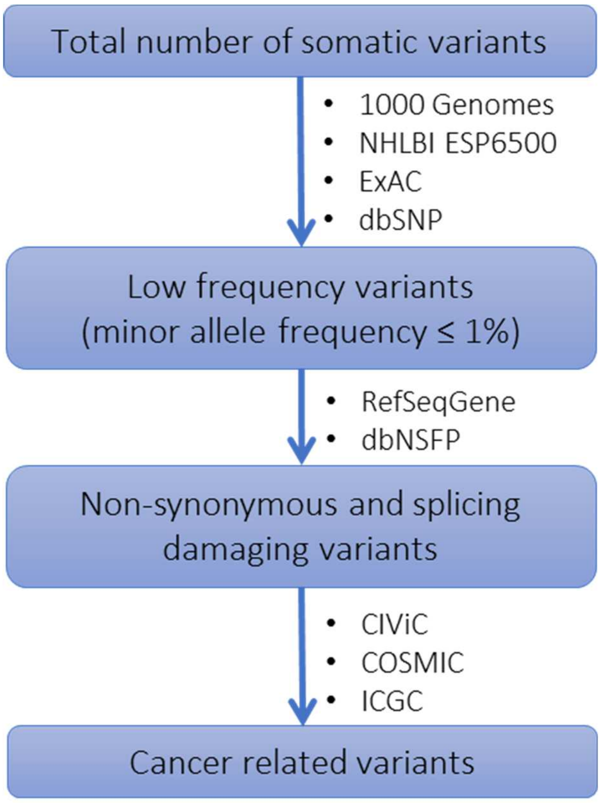


Figure 10. Variants annotation and filtering

With this workflow, we can select low frequency alterations to be evaluated at a deeper level by deciphering their biological significance. Moreover, the use of specific Cancer database enable the direct identification of mutations that inform targeted molecular therapies, drug sensitivity and prognosis for specific cancers.

IDENTIFICATION OF DRIVER GENES

The last step of the pipeline was aimed to the identification of driver carrier genes in the cohorts analysed based on somatic mutations identified. This step was performed applying a statistical analysis based on three distinct software (MutSigCV [27], OncodriveClust [28] and OncodriveFM [29]) using complementary and independent criteria aimed to detect positive selection signals. To maximize the sensitivity of driver genes detection step the results from the three methods were combined.

MutSigCV

To run the MutSigCV module three files were necessary:

- MAF mutation file: A Mutation Annotation Format (MAF) file is a tab-delimited text file that lists mutations.
- Coverage file: A tab-delimited file that gives the maximum number of bases covered to adequate depth in order to call mutations. The file allows MutSigCV to operate assuming full coverage.
- Covariates file: This file contains the genomic covariate data for each gene, for example, expression levels and DNA replication times, that will be used in MutSigCV to judge which genes are close to each other in mathematical "covariate space."

The vcf files of each tumoral-normal pairs were converted to the MAF file required by the software using vcf2maf-master and VEP. For the coverage and the covariates files were used the exome_full192.coverage.txt file and the gene.covariates.txt provided by the software.

```
#VCF CONVERSION
perl vcf2maf-master/vcf2maf.pl --vep-path VEP/ensembl-tools-release-
78/scripts/variant_effect_predictor/ --vep-data VEP/data/ --ref-fasta
VEP/data/homo_sapiens/78_GRCh37/Homo_sapiens.GRCh37.75.dna.primary_assembly.fa --
input-vcf $vcf --output-maf ${vcf}.maf --tumor-id $tumor --normal-id $normal

#MutSigCV

MutSigCV_1.4/MutSigCV_1.4/run_MutSigCV.sh mutations.maf exome_full192.coverage.txt
gene.covariates.txt mutsig mutation_type_dictionary_file.txt chr_files_hg19 1
```

OncodriveFM

To run OncodriveFM were necessary the files with the functional prediction for each tumoral-normal pair. The prediction uses were SIFT, Polyphen2 and MutationAssessor.

These files were prepared using ANNOVAR and converted to the format required by the software. For the mappings between genes and pathways to be analysed were used the file ensg_kegg.tsv provided by the software.

```
#FUNCTIONAL PREDICTION ANNOTATION
annovar/convert2annovar.pl -format vcf4 -allsample -withfreq $file >${file}.avinput;
annovar/annotate_variation.pl -filter -dbtype 1000g2014oct_all -buildver hg19 -maf 0.01 -out
${file} ${file}.avinput annovar/humandb/
annovar/table_annovar.pl ${file}.hg19_ALL.sites.2014_10_filtered annovar/humandb/ -buildver
hg19 -out ${file} -remove  --onetranscript -protocol ensGene,ljb26_all -operation g,f -nastring .


#OncodriveFM
oncodrivefm -e median -m ensg_kegg.tsv oncodrivefm.txt
```

OncodriveCLUST

To run OncodriveCLUST were necessary two separated list, one with the NON-Synonymous mutations file and one with the Synonymous mutations.

These files were prepared using the files produced with ANNOVAR [39] for OncodriveFM. Then were used several files provided by the software: CGC.phenotype.tsv that contains the Cancer Genome Consortium data; pfam_domains.txt that contains the gene domains and gene_transcrips.tsv that contains transcripts length for each gene.

```
#OncodriveCLUST
oncodriveclust -m 3 -c --cgc CGC_phenotype.tsv --dom pfam_domains.txt
oncodrivecluster_nonsyn.txt oncodriveclust_syn.txt gene_transcripts.tsv
```

RESULTS

PREPROCESSING RESULTS

The pipeline for WES analysis was applied to all the 205 sequenced leukemia samples. Each sample generated on average 61.7 million of fragments (100 nt X 2), and more than 93% of these data passed the QC filtering step, thus demonstrating the high quality of the generated data. The big majority of the filtered reads could then be mapped to the reference genome (80% on average). Moreover, aligned data showed a mean read depth of 86.5X and about 84% of the exome was represented at a minimum read depth of 10X, thus ensuring a highly comprehensive analysis of the whole exome. Detailed statistics of the total number of fragments reads, the total number of filtered and mapped fragments obtained for each sample are reported in Appendix 1. Detailed description on the average coverage after filtering and deduplication of the fragments and the percentage of target bases covered by at least 1, 10, 20 reads are reported in Appendix 2.

VARIANT CALLING RESULTS

The application of the variant calling pipeline enabled the identification of 8.208 somatic variants in AML patients and 5.582 in ALL patients with a mean per patient of 128 variants in AML and 151 in ALL. Of these, respectively 7.365 and 4.676 were unique, that means present only in a single patient.

Table 4 reports the summary statistics of variant calling on the two different leukemia types. The table includes statistics on the variants located in the CDS (coding DNA sequence) or in regions involved in mRNA splicing that may change the aminoacidic composition of the mRNA and thus affect the final protein product. In addition, the table reports the *loss of function/missense* variants, i.e. the most important candidate in driving the development of cancer.

		# Total	# CDS / splicing	# Loss of function / missense
AML 64 patients 128 samples	Unique	7.365	3.273	1.314
	Per patient (mean)	128	51	21
ALL 37 patients 77 samples	Unique	4.676	1.968	808
	Per patient (mean)	151	53	22

Table 4. Summary statistics of variant calling: total somatic variants, somatic variants located in CDS or splice sites, and Loss of function / missense somatic variants. The number of variants reported are either the total unique ones or the average per patient.

The bar graphs in Figure 11 and 13 show the total number of the somatic variants detected in each patient, with AML or ALL respectively. The pie charts in Figure 12 and 14 indicate the percentage of somatic variants divided according to their location in the gene or their potential effect on the encoded protein.

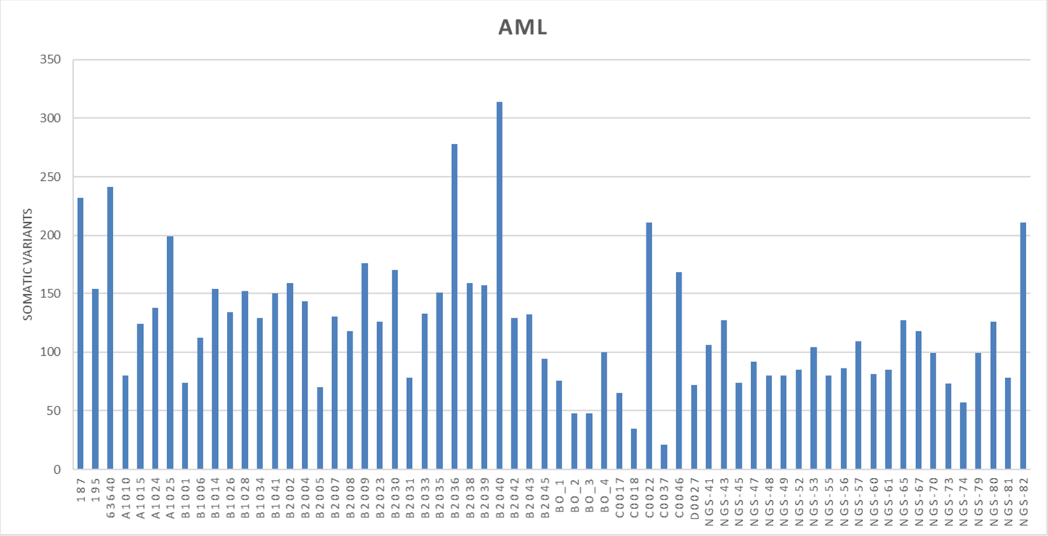


Figure 11. Distribution of somatic variants across the AML patients

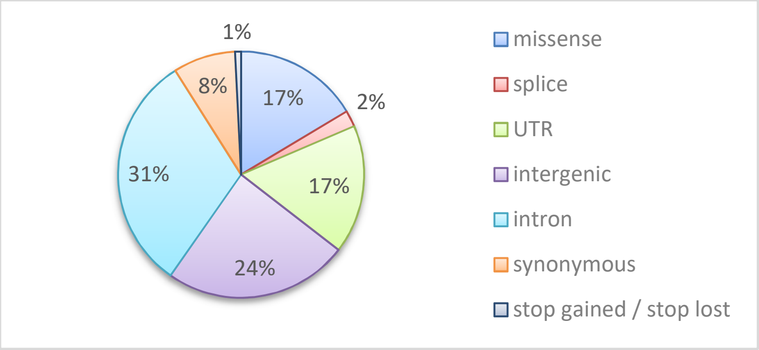


Figure 12. Distribution of somatic variants according to their putative functional effect in AML

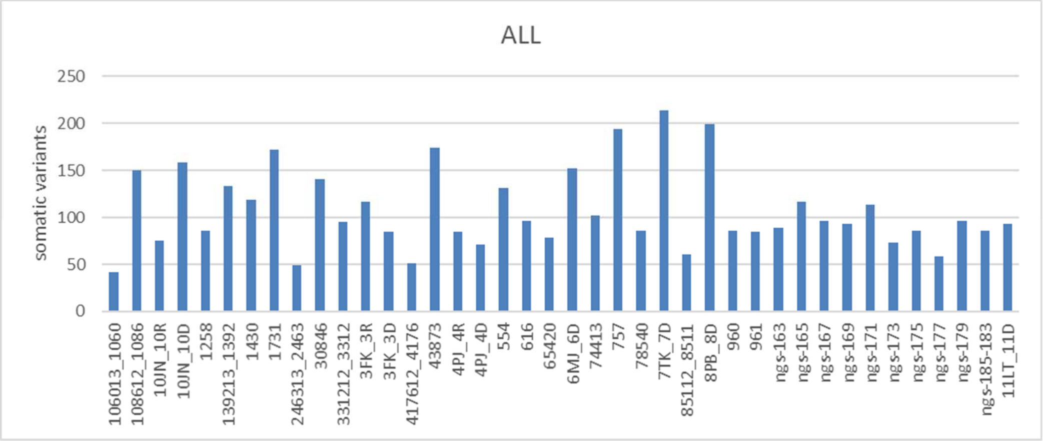


Figure 13. Distribution of somatic variants across the ALL patients

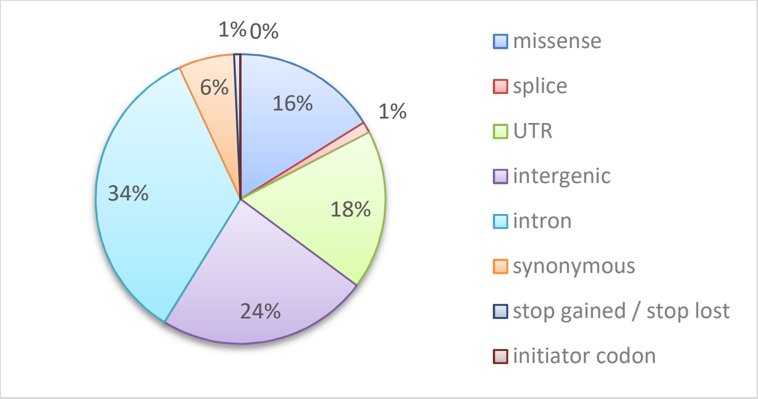


Figure 14. Distribution of somatic variants according to their putative functional effect in ALL

ANNOTATION RESULTS

The application of the variant annotation procedure enabled us to have a first insight into the genes that carry more mutations in the different leukemia patients (Figure 15-16) and to pone the basis for the application of the last and most important part of the pipeline, the identification of driver genes. The total number of mutated genes identified were 3.956 in AML and 2.821 in ALL.

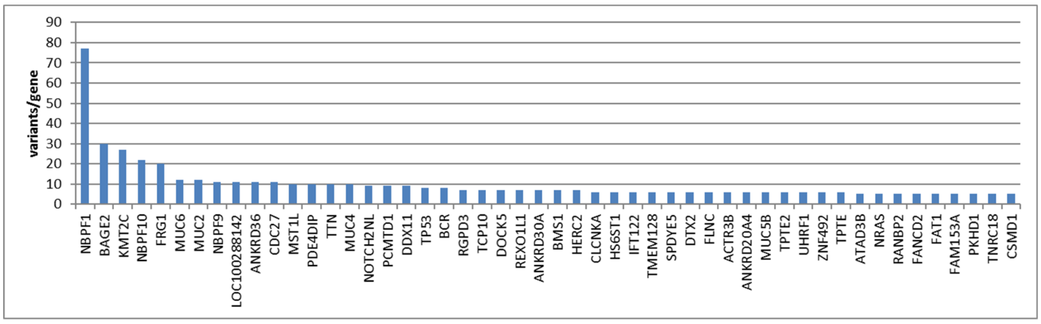


Figure 15. Top 50 mutated genes in AML samples

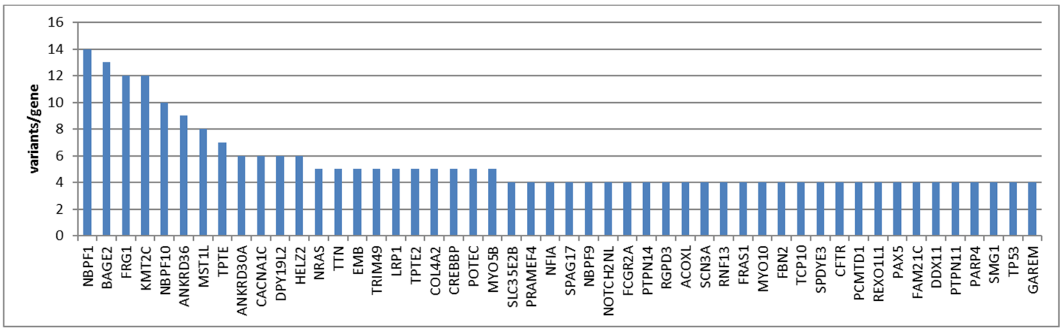


Figure 16. Top 50 mutated genes in ALL samples

As expected, some genes known to be involved in pathogenesis of leukemia and Cancer in general (i.e. TP53 and NRAS), were frequently mutated in the analysed samples, both in ALL and AML cohorts. However, some genes frequently mutated in the samples analysed are not associated with leukemia but are rather genes that accumulate more mutations in respect to the normal average rate of mutation (i.e. MUC2). To discriminate these types of mutations and identify the genes associated with leukemia it is necessary to apply a statistical analysis, as described in the subsequent application of tools for the discovery of driver candidate genes.

Moreover, to select the most-likely damaging variants a population frequency filter, based on the database ExAC, 1000 Genomes and NHLBI ESP6500, was used. A total of 5.871 variants in AML and 4.002 in ALL had a minor allele frequency (MAF) lower than 1% in the three selected population frequency databases, with 1.076 AML and 750 ALL being *loss of function* or *missense* variants.

Among the identified variants, some hundreds were previously annotated in the databases that contain variants identified by previous cancer studies (COSMIC, ICGC and CIViC), (Table 5).

		COSMIC	ICGC	CIViC
AML	Total	726	2.334	14
	Lof/missense MAF<1%	188	178	12
ALL	Total	283	1.154	2
	Lof/missense MAF<1%	118	118	2

Table 5. Identified Variants annotated in Cancer related databases: total number of variants and total number of Loss of function / missense variants with a minor allele frequency lower than 1% in the population frequency databases

The figures below report the distributions of the identified variants that were present in the COSMIC (Figure 17-18) and in ICGC (Figure 19-20) databases, grouped by the origin of cancer (organ or tissue) where they found by the original study. These figures show that in both AML and ALL samples a huge number of variants were annotated in Haematopoietic and Lymphoid tissue in COSMIC and in the blood tissue in ICGC. A summary of these variants is reported in Table 6.

		COSMIC	ICGC
AML	Total	90	655
	Lof/missense MAF<1%	34	37
ALL	Total	28	322
	Lof/missense MAF<1%	16	19

Table 6. Somatic variants identified by the study and annotated in Haematopoietic and Lymphoid tissue in or in Blood tissue in the COSMIC or ICGC database, respectively. Variants reported are the total number of variants and the total number of Loss of function / missense variants with a minor allele frequency lower than 1% in the population frequency databases

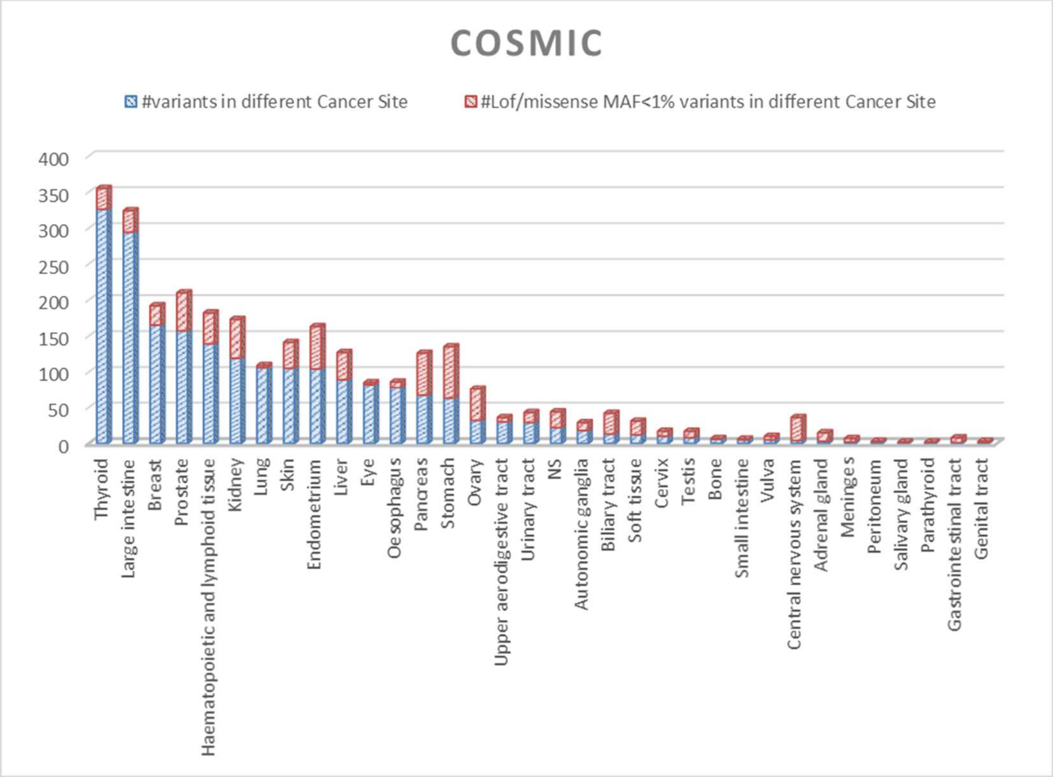


Figure 17. AML variants reported in COSMIC, divided by cancer origin

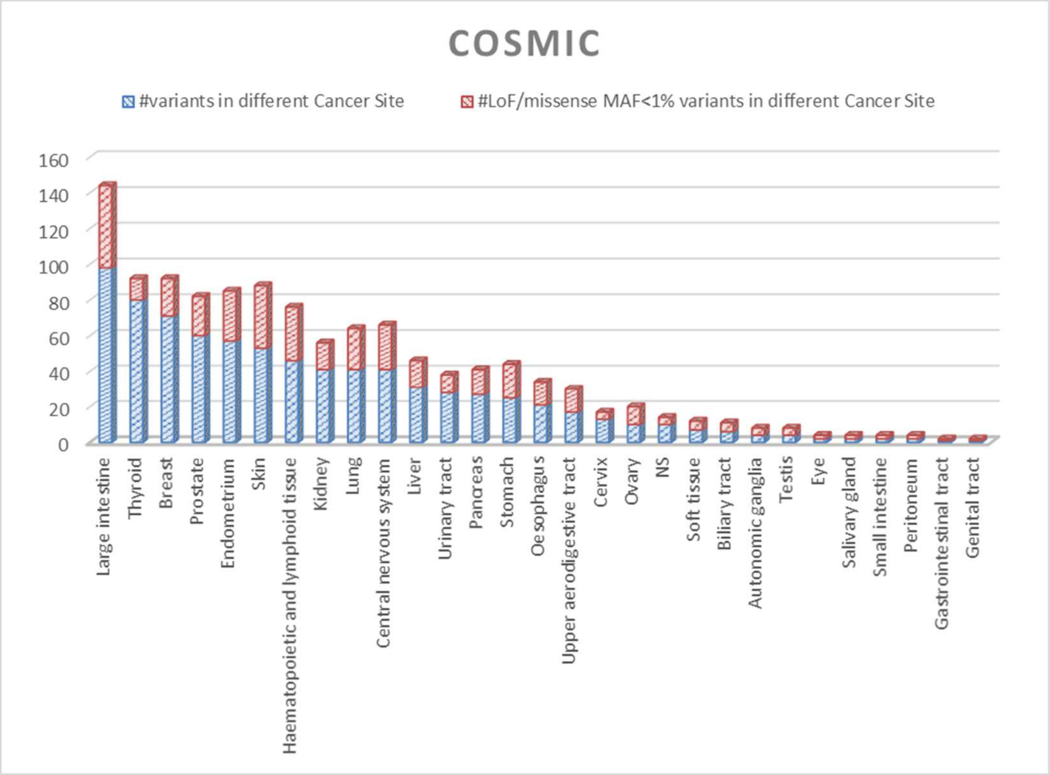


Figure 18. ALL variants reported in COSMIC, divided by cancer origin

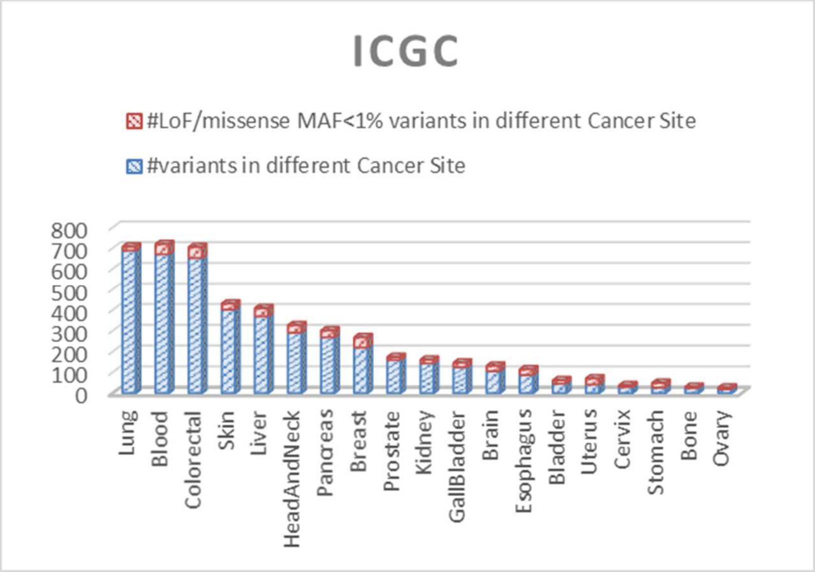


Figure 19. AML variants reported in ICGC, divided by Cancer origin

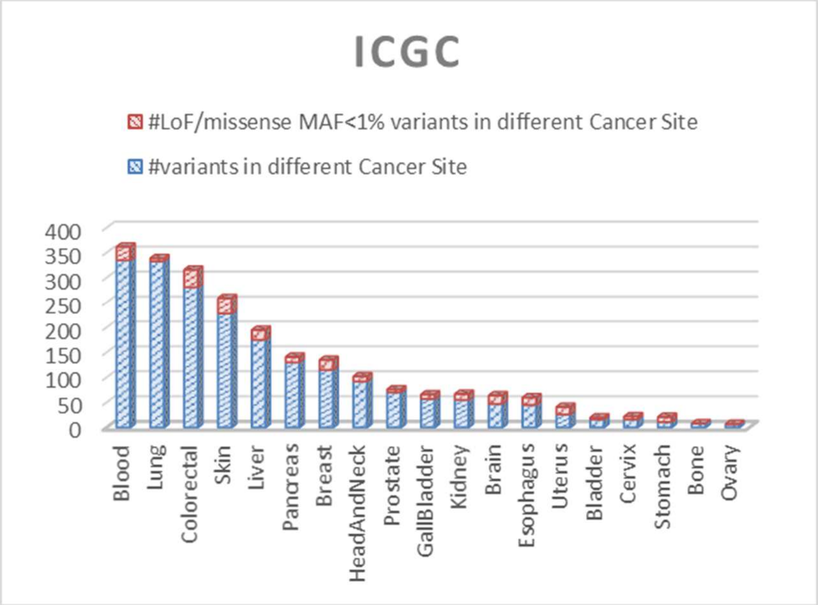


Figure 20. ALL variants reported in ICGC, divided by Cancer origin

The CIViC resource was interrogated to identify variants, among those retrieved in our analysis, that have been previously associated with good/bad response to a certain therapy or with a specific cancer outcome. Table 7 and 8 report the list of such variant.

Chr:Pos	Ref/Alt	Identifier	Gene Name	Disease	Drugs	Evidence Type	Clinical Significance
1:115256529	T/C	rs11554290	NRAS	Melanoma	Temozolomide	Predictive	Sensitivity
1:115258744	C/T	rs121434596	NRAS	Melanoma	17-AAG	Predictive	Sensitivity
2:198266834	T/C		SF3B1	Breast Cancer	Spliceostatin A	Predictive	Sensitivity
2:209113112	C/T	rs121913500	IDH1	Anaplastic Oligodendroglioma	AG-5198	Predictive	Sensitivity
2:209113113	G/A	rs121913499; rs121913501	IDH1	Acute Myeloid Leukemia	GSK321	Diagnostic,Prognostic,Predictive	Positive,N/A,Sensitivity
4:55599321	A/T	rs121913507	KIT	Acute Myeloid Leukemia,Systemic Mastocytosis	Midostaurin	Prognostic,Predictive	Poor Outcome,Sensitivity, Poor Outcome
9:21975017	C/T	rs3814960	CDKN2A	Esophagus Squamous Cell Carcinoma		Prognostic	Poor Outcome
12:25398281	C/T	rs112445441	KRAS	Colorectal Cancer	Cetuximab	Predictive	Sensitivity,Sensitivity,Resistance or Non-Response,Resistance or Non-Response,Resistance or Non-Response

12:25398284	C/G	rs121913529;
		rs121913531;
		rs121913534
			KRAS	Lung Adenocarcinoma	Gefitinib,Erlotinib	Predictive	Resistance or Non-Response

12:25398284	C/T	rs121913529;
		rs121913531;
		rs121913534
			KRAS	Hairy Cell Leukemia,Lung Cancer,Non-small Cell Lung Carcinoma,Pancreatic Carcinoma,Colorectal Cancer,Pancreatic Cancer,Tumor Of Exocrine Pancreas,Pancreatic Ductal Carcinoma	ARRY-142886,BEZ235 (NVP-BEZ235,Dactolisib),MK-2206,Cetuximab,Vemurafenib	Diagnostic,Predictive,Prognostic	Positive,Sensitivity,Sensitivity,Sensitivity, Poor Outcome,Poor Outcome,Resistance or Non-Response,Poor Outcome

12:25398285	C/A	rs121913530	KRAS	Lung Cancer,Non-small Cell Lung Carcinoma,Cancer,Colorectal Cancer,Non-small Cell Lung Carcinoma	Selumetinib (AZD6244),Docetaxel,ARS-853,EGFR Inhibitor,Gefitinib,Erlotinib	Diagnostic,Predictive,Prognostic	Positive,Sensitivity,Sensitivity,Resistance or Non-Response,Poor Outcome
12:111884608	T/C	rs3184504	SH2B3	Colorectal Cancer		Predisposing	Positive
15:90631838	C/T	rs121913503	IDH2	Acute Myeloid Leukemia,Myelodysplastic Syndrome		Prognostic	Poor Outcome
17:7577538	C/T	rs11540652	TP53	Breast Cancer		Prognostic	Poor Outcome

Table 7. AML variants reported in the CIViC database. Chr:Pos, chromosome and position in the genome; Ref/Alt, reference and alternative alleles; Gene Name, name of the gene where the variant reside; Disease, phenotype associated to the variant; Drugs, treatment evidence; Evidence Type, the predictive / prognostic / diagnostic association between an evidence statement and a variant; Clinical significance, the sub-type of evidence type that the statement presents.

Chr:Pos	Ref/Alt	Identifier	Gene Name	Disease	Drugs	Evidence Type	Clinical Significance
1:115258744	C/T	rs121434596	NRAS	Melanoma	17-AAG	Predictive	Sensitivity
12:25398284	C/T	rs121913529;rs121913531;rs121913534	KRAS	Hairy Cell Leukemia,Lung Cancer,Non-small Cell Lung Carcinoma,Pancreatic Carcinoma,Colorectal Cancer,Pancreatic Cancer,Tumor Of Exocrine Pancreas,Pancreatic Ductal Carcinoma	ARRY-142886,BEZ235 (NVP-BEZ235,Dactolisib), MK-2206,Cetuximab,Vemurafenib	Diagnostic,Predictive,Predictive,Predictive,Prognostic,Prognostic,Predictive, Prognostic	Positive,Sensitivity,Sensitivity,Sensitivity, Poor Outcome,Poor Outcome,Resistance or Non-Response,Poor Outcome

Table 8. ALL variants reported in the CIViC database. Chr:Pos, chromosome and position in the genome; Ref/Alt, reference and alternative alleles; Gene Name, name of the gene where the variant reside; Disease, phenotype associated to the variant; Drugs, treatment evidence; Evidence Type, the predictive / prognostic / diagnostic association between an evidence statement and a variant; Clinical significance, the sub-type of evidence type that the statement presents.

Among the variants identified in AML patients, three variants were already associated with the disease by previous studies (Table 9). These variants are related to the specific diagnosis and prognosis of the disease, and one of them is associated with the response to a specific drug, i.e. Midostaurin that in a phase II clinical trial shows that 60% of patients (N=89) responded to treatment.

Chr:Pos	Ref/Alt	Identifier	Gene Name	Disease	Drugs	Evidence Type	Clinical Significance	Evidence Statement
2:2091131	G/A	rs121913499; rs121913501	IDH1	Acute Myeloid Leukemia	GSK321	Diagnostic, Prognostic, Predictive	Positive, N/A, Sensitivity	IDH1 R132 mutation is associated with patients of older age, high platelet count during diagnosis, cytogenic normalcy and NPM1 mutation., IDH1 R132 mutation in patients with AML is not associated with any prognostic value compared to patients with wild-type IDH1.,Newly developed allosteric inhibitors (GSK321) of IDH1 led to granulocytic differentiation in-vitro and in-vivo.
4:5559932	A/T	rs121913507	KIT	Acute Myeloid Leukemia	Midostaurin	Prognostic	Poor Outcome	In acute myloid leukemia patients, D816 mutation is associated with earlier relapse and poorer prognosis than wildtype KIT.
15:906318	C/T	rs121913503	IDH2	Acute Myeloid Leukemia		Prognostic, Prognostic	N/A, Poor Outcome	AML patients with IDH2 mutations such as R172K have event free survival and overall survival similar to those with wild-type IDH2.,In AML, patients with an IDH2 R172K mutation have worse overall survival compared to those with wild-type IDH2.

Table 9. AML variants reported in CIVIC already associated with AML

IDENTIFICATION OF DRIVER GENES

To identify genes carrying driver somatic mutations, we employed three statistical tools, namely MutSigCV, OncodriveFM, OncodriveCLUST, on the sets of annotated mutations for each leukemia type. The identified genes with signals of positive selection were then mapped into an interaction network using Cytoscape 3 Reactome FI plugin. Gene modules of the interaction network were identified through a clustering approach and the most most significant markers within such modules were identified by performing an enrichment analysis to identify pathways involved in the tumorigenesis.

A total of 64 AML patients were analysed with the selected software and 17 genes with signals of positive selection were identified as potential driver carriers by at least one bioinformatic approach (Table 10).

Gene	non-synonymous mutations	patient(s)	MutSigCV Recurrence	OncodriveCLUST Clustering	OncodriveFM Functional Impact
AGGF1	3	3		**X**	
CDC27	13	11		**X**	
DPY19L2	5	5		**X**	
FRG1	14	12	**X**		
FRG2B	3	1		**X**	
H2AFV	7	6	**X**	**X**	
IDH1	3	3		**X**	
IDH2	3	3		**X**	
KRAS	7	7	**X**	**X**	
KRT8	3	2		**X**	
MUC6	14	8		**X**	
NRAS	4	3		**X**	
PHGR1	3	2		**X**	
RGPD3	14	9		**X**	
SEC63	4	3		**X**	
SF3B1	5	5		**X**	
SMC1A	3	3		**X**	

Table 10. List of potential driver carriers genes identified by the statistical methods, with the total number of mutations and patients carrying a mutation on the indicated gene

Of the total 17 genes identified, 14 were mapped in the functional interaction network with 14 linker genes. Clustering of these genes identified six modules in the network (Table 11, Figure 21).

Module	Nodes in Module	Node List
0	7	GRB2,KRAS,KRT8,NRAS,PPP2CA,SOCS3,YWHAQ
1	6	CDC27,H2AFV,HIST1H2BA,RPS27A,SEC61A2,SEC63
2	4	CWC22,FRG1,SF3B1,SMC1A
3	4	IDH1,IDH2,PC,PSMD12
4	3	AGGF1,FOS,RBPJ
5	2	MUC6,TFF1

Table 11. The six modules identified in the interaction network of the potential driver carriers genes in the AML patients

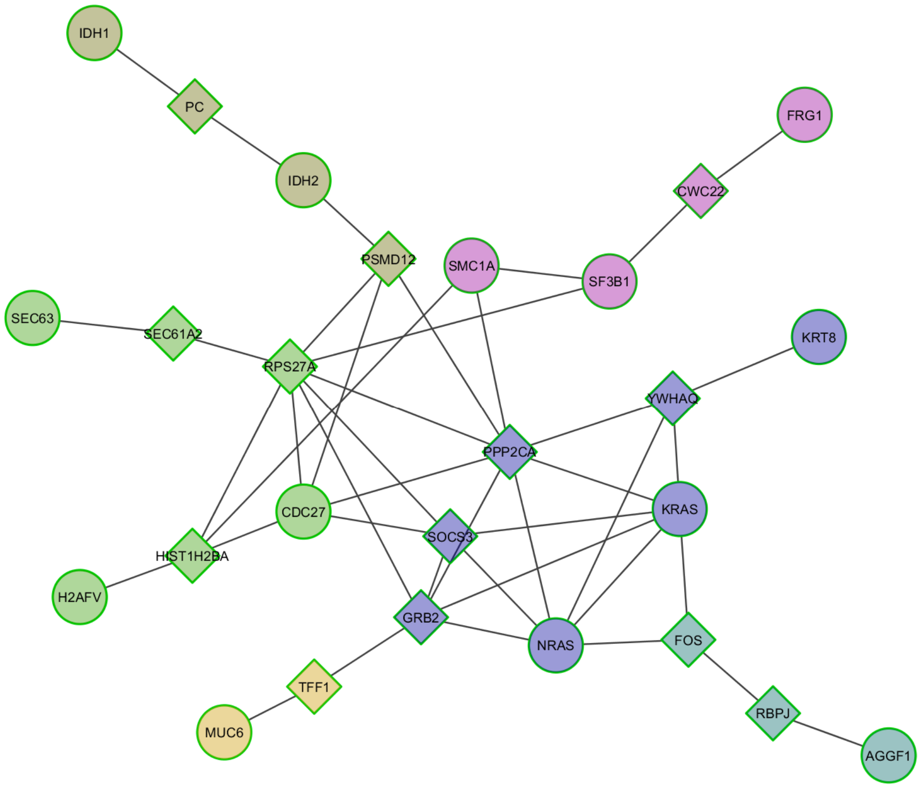


Figure 21. Interaction network of the potential driver carriers genes in the AML patients and the six modules identified (each indicated with a different colour).

The methods applied have identified known leukemia pathways, like the NRAS/KRAS (Table 12) and IDH1/IDH2 (Table 13) interaction modules, as significantly enriched (FDR-adjusted p-value < 0.05) in the network, thus demonstrating the validity of the approach.

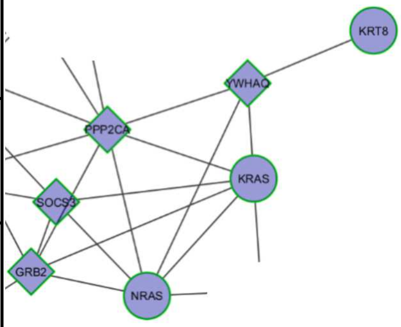
Module	GeneSet	FDR	Nodes	
1	RAF/MAP kinase cascade(R)	1.00E-03	NRAS,KRAS	
1	Ras signaling in the CD4+ TCR pathway(N)	1.00E-03	NRAS,KRAS	
1	Signaling by Leptin(R)	1.33E-03	NRAS,KRAS	
1	p53 pathway feedback loops 2(P)	1.25E-03	NRAS,KRAS	
1	EGF receptor (ErbB1) signaling pathway(N)	1.00E-03	NRAS,KRAS	

Table 12. NRAS/KRAS module enriched in the AML patient

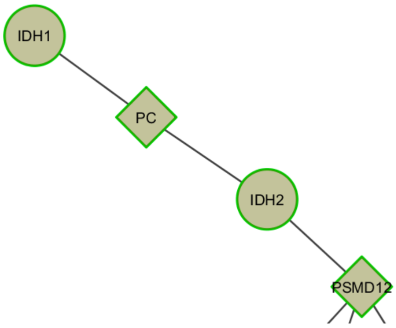
Module	GeneSet	FDR	Nodes	
4	2-Oxocarboxylic acid metabolism(K)	<1.000e-03	IDH2,IDH1	
4	Citrate cycle (TCA cycle)(K)	<5.000e-04	IDH2,IDH1	
4	Glutathione metabolism(K)	6.67E-04	IDH2,IDH1	
4	Biosynthesis of amino acids(K)	7.50E-04	IDH2,IDH1	
4	Peroxisome(K)	1.20E-03	IDH2,IDH1	
4	Carbon metabolism(K)	1.33E-03	IDH2,IDH1	
4	TCA cycle(P)	9.71E-03	IDH2	
4	Peroxisomal lipid metabolism(R)	6.99E-02	IDH1	

Table 13. IDH1/IDH2 module enriched in the AML patient.

Statistical analysis of the 38 patients affected by ALL identified 29 genes with signals of positive selection as potential carriers of driver mutations (Table 14).

Gene	# non-synonymous mutations	# patient(s)	MUTSIG Recurrence	ONCODRIVECLUST Clustering	ONCODRIVEFM Functional Impact
AGAP10	3	3		X	
ANK3	4	4		X	
ANKS1B	5	3		X	
CCDC83	4	4		X	
CFHR1	4	2		X	
CS	6	3	X		
DDN	4	4		X	
DSPP	3	2		X	
EBPL	3	2		X	
H2AFV	4	2		X	
JAK2	3	2		X	
KIF9	1	1		X	
KRAS	4	4		X	
LRP1B	3	3		X	
MUC20	9	2		X	
MYH7	3	3		X	X
NRAS	13	12	X	X	X
PAX5	6	6	X	X	X
PDIA4	3	3		X	
PGM1	4	2		X	
PHKG1	3	2		X	
PRKRIR	6	3	X	X	
RGPD3	9	5		X	
SEC63	3	3		X	
SIRT4	2	2			X
TMEM147	1	1		X	
TP53	4	4			X
TTC7B	3	3		X	
ZP3	4	4		X	

Table 14. List of potential driver carriers genes selected by the statistical methods, with the total number of mutations and patients involved

Of the 29 genes identified, 20 were mapped in the functional interaction network with 21 linker genes. Clustering identified seven enriched modules in the network (Table15, Figure 22).

Module	Nodes in Module	Node List
0	11	B4GALT1,EP300,H2AFV,HDAC2,KIF9,PAX5,PRKRIR,SIN3A,STK4,TP53,ZP3
1	11	ANK3,EGFR,GRB2,IL2RG,JAK2,KRAS,MUC20,NRAS,SFN,SOS1,SPTB
2	8	C1R,CALM1,CFHR1,JUN,MYH7,PAFAH1B1,PDIA4,PHKG1
3	4	CS,FDPS,MDH2,PGM1
4	3	RPS27A,SEC61A2,SEC63
5	2	APBB2,LRP1B
6	2	DSPP,ITGB1

Table 15. The seven modules identified in the interaction network of the potential driver carriers genes in the ALL patients

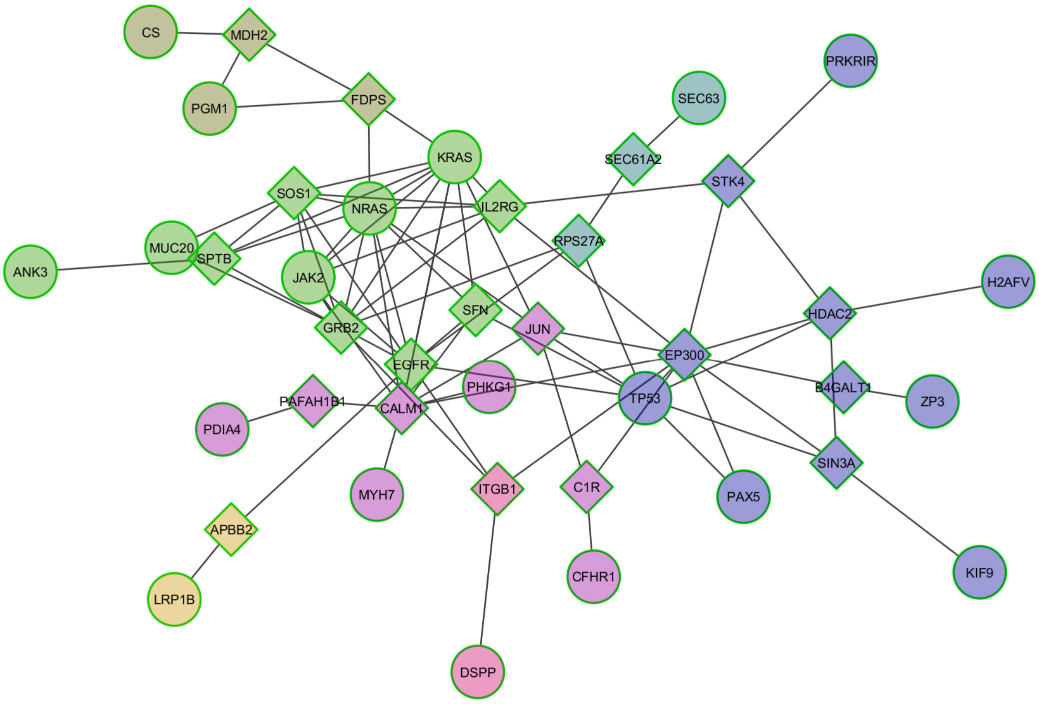


Figure 22. Interaction network of the potential driver carriers genes in the ALL patients and the seven modules identified

Also in the case of ALL patients, the methods applied identified two known leukemia pathways as significantly enriched: the TP53 (table 16) and NRAS/KRAS/JAK2 (Table 17) interaction modules.

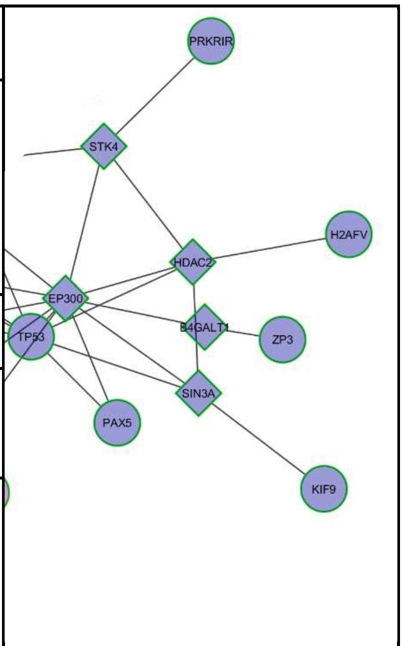
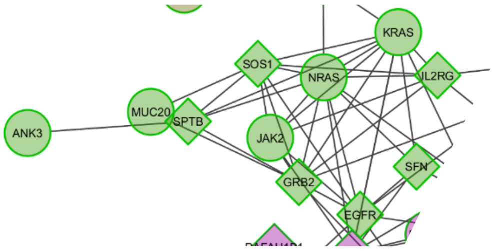
Module	GeneSet	FDR	Nodes	 PRKRIR STK4 H2AFV HDAC2 EP300 B4GALT1 ZP3 TP53 SIN3A PAX5 KIF9
1	Factors involved in megakaryocyte development and platelet production(R)	2.30E-01	TP53,KIF9	
1	PLK3 signaling events(N)	2.16E-01	TP53	
1	P53 pathway feedback loops 1(P)	1.75E-01	TP53	
1	Transcriptional misregulation in cancer(K)	1.38E-01	TP53,PAX5	

Table 16. TP53 module enriched in the ALL patient

Module	GeneSet	FDR	Nodes	 KRAS SOS1 NRAS IL2RG MUC20 SPTB ANK3 JAK2 SFN GRB2 EGFR
2	Signaling by Leptin(R)	<1.000e-03	NRAS,KRAS,JAK2	
2	GMCSF-mediated signaling events(N)	<3.333e-04	NRAS,KRAS,JAK2	
2	ErbB2/ErbB3 signaling events(N)	<3.333e-04	NRAS,KRAS,JAK2	
2	Interleukin-2 signaling(R)	<2.500e-04	NRAS,KRAS,JAK2	
2	SHP2 signaling(N)	<2.000e-04	NRAS,KRAS,JAK2	
2	Prolactin signaling pathway(K)	<1.667e-04	NRAS,KRAS,JAK2	

2	RAF/MAP kinase cascade(R)	<1.429e-04	NRAS,KRAS	
2	Cholinergic synapse(K)	<1.250e-04	NRAS,KRAS,JAK2	
2	PDGFR-beta signaling pathway(N)	<1.111e-04	NRAS,KRAS,JAK2	
2	Ras signaling in the CD4+ TCR pathway(N)	<1.000e-04	NRAS,KRAS	

Table 17. NRAS/KRAS/JAK2 module enriched in the ALL patient

The statistical analysis led to the identification of a total of 32 markers (including globally 19 novel and 9 established ones) across these leukemia types as reported in Table 18.

Leukemia type	**Genes identified**	
	Novel genes	**Established genes**
AML	H2AFV, SEC63, SMC1A, AGGF1, CDC27, FRG1	IDH1, IDH2, KRAS, NRAS, SF3B1,
ALL	TMEM147, TTC7B, ANK3, CFHR1, CS, H2AFV, KIF9, PHKG1, PRKRIR, SEC63, SIRT4, PGM1, RGPD3, DDN, LRP1B	TP53, JAK2, KRAS, NRAS, PAX5, ANKS1B

Table 18. Gene markers selected on statistical and network-based analysis.

DISCUSSION

In the last 10 years, NGS technology became a trustworthy method to study diseases with a genetic basis. By enabling the discovery of disease-associated mutations, NGS provides the foundation for a wide range of applications in translational research (i.e. Cancer studies).

The aim of the project presented was the application of WES analysis to patients affected by leukemia, either AML or ALL, to uncover their genetic variability and to find new markers to help the diagnosis and identify the prognosis of these malignancies. In this context, the work conducted focused on the setup and application of a bioinformatic pipeline that allows the identification of the somatic variants carried by each patient, their correlation with the available knowledge in the Cancer Genomics area and the identification of markers for AML and ALL leukemia. Given that the distinction between *"driver"* mutations, responsible for leukemia development, and *"passenger"* mutations is one of the greatest challenges in the field, one main goal of the present project was the application of dedicated statistics and bioinformatics strategies for the selection of the most relevant mutations.

The setup of a bioinformatic pipeline that enables the identification of a reliable set of somatic mutations has required the selection of tools suitable for the analysis of NGS data derived from cancer samples. The selection of dedicated software to perform the initial pre-processing of the data, like removing some known errors due to technological bias, guarantees the use of sequencing data of high quality and ensures that the subsequent analysis will be performed on well-generated data. This is of utmost importance when considering that the majority of variants identified occurred in only a subset of the fragments analysed, therefore the starting data must be as clean as possible from additional confounding variables. Similarly, the selection of a variant caller suitable to detect low frequency variants, that represent the cancer sample, has been crucial to overcome specific problems related to the heterogeneity nature of cancer samples. Thus, the application of MuTect allowed the identification of a large and reliable set of somatic variants to be evaluated for the identification of new biomarkers and driver genes. Overall, the selection of the most suitable bioinformatic pipeline and its application on all the sequenced leukemia samples has required a substantial amount of time but has assured the generation of high quality data, as demonstrated by the big number of sequenced fragments that passed the QC filtering step and the good exome coverage obtained. Subsequently, the application of the variant calling pipeline has enabled the identification of a huge number of somatic variants, and the further selection of meaningful variants, *e.g.* with a potential impact on the gene product, previously associated to cancer development or enriched in driver genes. Moreover, among all the variants identified, 4291 variants in AML and 3237 in ALL were never associated to cancer previously, thus representing a good starting point for the discovery of novel biomarker.

The correlation of the identified somatic variants with the biological knowledge present in different databases allowed to identify the variants most-likely responsible of leukemia development (driver mutations).

The first database utilized at this aim was RefSeqGene that enabled us to correctly identify the protein-coding genes in which the variant resides and to assess its functional consequence on the protein product, i.e. location within the CDS or on splice regions, and among these the *loss of function/missense* variants. In addition, RefSeqGene allowed us to have a first insight into the genes most frequently mutated in the different leukemia patients and to pone the basis for the identification of driver genes. The results obtained were reliable as demonstrated by the identification of genes that have been already associated to cancer pathogenesis (i.e. TP53, NRAS). Most importantly, our results also highlighted other genes that are frequently mutated in leukemia and that were never associated to this type of cancer before, these were 19 in total and included for example CDC27 and LRP1B.

To further narrow down the list of relevant somatic variants, we selected those that: (i) were rare, i.e. had a low frequency in healthy reference populations, (ii) were annotated in databases collecting variants associated to cancer by previous studies, (iii) were enriched in driver genes as identified by selected statistical methods.

Selection for frequency allowed to filter out innocuous common variants, thus decreasing the total number of potentially dangerous variants from 8.208 to 5.871 and from 5.582 to 4.002, respectively for AML and ALL. Further merging of these data with resources that contain variants coming from previous cancer studies, highlighted that a big number of variants were already associated to cancer of Haematopoietic and Lymphoid tissue (90 in AML and 28 in ALL) and blood (655 in AML and 322 in ALL), indicating that the selected somatic mutations can have an impact on the tissues involved in leukemia development. Moreover, interrogating the CIViC resource, among the variants identified in AML patients, three were already associated with the disease by previous studies, two of these already related to a poor prognosis and one of them was associated with the good response to a specific drug (i.e. Midostaurin).

Overall, only with the application of the right biological knowledge we can obtain information of fundamental importance in the analysis of single leukemic patients, enabling the application of a specific tailored therapy selected on the basis of mutations carried by each patient. However, resources connecting mutations to good/bad response or prognosis are still not complete. Still, they can take great vantage of large sequencing project like the one presented here to obtain novel biomarkers that can be further validated and then used for addressing the most appropriate therapy on newly diagnosed patients. Therefore, the last part of the project was dedicated to the identification of genes that are most likely implicated in the development of the disease. In fact, as in the cancer genome only a small subset of the somatic mutations found in the cells are responsible for tumorigenesis we discriminated between real driver mutations from passenger mutations, by identifying the genes that exhibits signals of positive selection across our cohort of tumour samples. To perform this task, we employed three statistical tools that together allowed us to obtain the most comprehensive list of driver genes, overcoming the intrinsic limitation of each software taken individually. This analysis led to the identification of a total of 32 potential biomarkers (including 19 novel and 9 established ones) across all the samples. Subsequent enrichment analysis highlighted the genes involved in the tumorigenesis and demonstrated the significance of the markers identified. We identified pathways known to be implicated in leukemia development, like the NRAS/KRAS and IDH1/IDH2 modules in AML, and the TP53 and NRAS/KRAS/JAK2 modules in ALL. Beside these, the analysis found enriched pathways that are not connected with leukemia in an established manner. These include interesting relevant candidates that can be involved in leukemia pathogenesis: CDC27 or Cell division cycle 27, is a protein involved in the regulation of the cell cycle, interesting in our condition because the dysregulated cell cycle progression has a critical role in tumorigenesis/leukemia. Indeed, in colorectal cancer CDC27 expression is significantly correlated with tumor progression and poor patient survival [40]; LRP1B or LDL receptor related protein 1B is a gene that encodes a member of the low density lipoprotein (LDL) receptor family. These receptors play a wide variety of roles in normal cell function and development due to their interactions with multiple ligands. LRP1B point mutations have been reported in a significant percentages of lung cancer [41] as well as in melanoma [42] and triple negative breast cancer [43]. One of the novel gene identified has a specific role in the activation of the immune system: PRKRIR is a protein-kinase that enhances the antiviral response, a crucial activity of lymphocytes [44]; even if its role in cancer is not well established, PRKRIR constitutes a promising candidate linking leukocyte dysregulation with cancer development. ANK3, ankyrin 3, is significantly mutated in endometrial cancer and in melanoma; it encodes for a membrane protein that play key roles in activities such as cell motility, activation, proliferation, contact and the maintenance of specialized membrane domains; these are important aspects in leukocyte biology, however the role of this gene is still not well established in the immune system yet (http://www.tumorportal.org/ANK3). Even if potentially relevant, the function of other genes identified has not been clearly connected with leukocyte biology or cancer development yet. Additional validation and functional studies will be necessary to investigate the implication of all the driver genes identified with leukemia pathogenesis and to define their role as potential biomarkers for disease prognosis and therapy response.

In conclusion, the study demonstrated that the application of NGS, in combination with an appropriate analysis pipeline and integration of a-priori biological knowledge can lead to the discovery of novel candidate biomarkers associated with leukemia development. This Proof-Of-Concept study demonstrated that the NGS approach has the potential to be applied routinely in the clinic to obtain crucial unprecedented information for an accurate and quick diagnosis and to guide tailored interventions on these malignancies, thus leading to great successful improvements in this field.

APPENDIX

APPENDIX 1

Detailed statistics on the total number of fragments, the total number of filtered and mapped fragments.

SAMPLE	Type of Leukemia	Phase	# sequenced fragments	# filtered fragments	% filtered fragments	# mapped fragments (dedup)	% mapped fragments (dedup)
Sample_187	AML	diagnosis	86029887	80836959	93,96%	47000322	58,14%
Sample_197	AML	germline	78469088	73800314	94,05%	42710753,5	57,87%
Sample_195	AML	diagnosis	78449806	74032573	94,37%	41817760	56,49%
Sample_198	AML	germline	93325767	87708025	93,98%	48686117	55,51%
Sample_63640	AML	diagnosis	87484829	82236643	94,00%	50358970,5	61,24%
Sample_199	AML	germline	33193541	31281332	94,24%	18075269,5	57,78%
Sample_A1010D	AML	diagnosis	69086706	65314721	94,54%	56729435,5	86,86%
Sample_A1010S	AML	germline	86645093	82100017	94,75%	71832845,5	87,49%
Sample_A1015Dbis	AML	diagnosis	53118847	49264808	92,74%	43714933	88,73%
Sample_A1015S	AML	germline	22498594	21336279	94,83%	19535567,5	91,56%
Sample_A1024D	AML	diagnosis	83830187	77226104	92,12%	68442084,5	88,63%
Sample_A1024S	AML	germline	29505251	27385046	92,81%	24930547,5	91,04%
Sample_A1025D	AML	diagnosis	82814040	77031343	93,02%	67315045	87,39%
Sample_A1025S	AML	germline	37050554	34810659	93,95%	31681783	91,01%
Sample_B1001D	AML	diagnosis	93440778	88440204	94,65%	78646235,5	88,93%
Sample_B1001S	AML	germline	69010621	65256891	94,56%	56095424	85,96%
Sample_B1006D	AML	diagnosis	57177046	53442497	93,47%	49119093,5	91,91%
Sample_B1006S	AML	germline	52609965	48698395	92,56%	44955797	92,31%
Sample_B1014D	AML	diagnosis	60160493	56320811	93,62%	47388697,5	84,14%
Sample_B1014S	AML	germline	23647544	22551048	95,36%	20222256,5	89,67%
Sample_B1026D	AML	diagnosis	53242686	49223059	92,45%	44783359,5	90,98%
Sample_B1026S	AML	germline	30701851	28937816	94,25%	26493257	91,55%
Sample_B1028D	AML	diagnosis	67679596	62679188	92,61%	57695404,5	92,05%
Sample_B1028S	AML	germline	32019454	29889207	93,35%	26978421	90,26%
Sample_B1034D	AML	diagnosis	46850219	44093577	94,12%	39288379	89,10%
Sample_B1034S	AML	germline	33523945	31590080	94,23%	28797169,5	91,16%
Sample_B1041D	AML	diagnosis	71839118	67324385	93,72%	45118157,5	67,02%
Sample_B1041S	AML	germline	32848233	31187890	94,95%	21765124	69,79%
Sample_B2002D	AML	diagnosis	61134067	56784815	92,89%	49584603	87,32%
Sample_B2002S	AML	germline	19738373	18729571	94,89%	17008451,5	90,81%
Sample_B2004D	AML	diagnosis	56088744	52917970	94,35%	49167606	92,91%
Sample_B2004S	AML	germline	28111907	26946179	95,85%	24873949,5	92,31%
Sample_B2005D	AML	diagnosis	57235519	53642391	93,72%	49972322,5	93,16%

Sample_B2005S	AML	germline	63612873	59342554	93,29%	55303446	93,19%
Sample_B2007D	AML	diagnosis	69080771	64543316	93,43%	57582479	89,22%
Sample_B2007S	AML	germline	48599102	45844183	94,33%	42235665	92,13%
Sample_B2008D	AML	diagnosis	105032896	99333952	94,57%	87617743,5	88,21%
Sample_B2008S	AML	germline	73277132	69388497	94,69%	58717822	84,62%
Sample_B2009D	AML	diagnosis	58801391	54970978	93,49%	51158965,5	93,07%
Sample_B2009S	AML	germline	25494454	24389015	95,66%	21772026,5	89,27%
Sample_B2023D	AML	diagnosis	54544548	49510634	90,77%	43662810,5	88,19%
Sample_B2023S	AML	germline	38761333	36047420	93,00%	33041307	91,66%
Sample_B2030D	AML	diagnosis	75680896	71876934	94,97%	65051977,5	90,50%
Sample_B2030S	AML	germline	30758899	28874272	93,87%	26602684	92,13%
Sample_B2031D	AML	diagnosis	45091853	42613721	94,50%	34516637,5	81,00%
Sample_B2031S	AML	germline	33463869	31214478	93,28%	28633463,5	91,73%
Sample_B2033D	AML	diagnosis	58109645	55018734	94,68%	48513546	88,18%
Sample_B2033S	AML	germline	29706329	27859919	93,78%	25190052,5	90,42%
Sample_B2035D	AML	diagnosis	58664146	53482322	91,17%	25901922	48,43%
Sample_B2035S	AML	germline	26572499	22463899	84,54%	14537532,5	64,72%
Sample_B2036D	AML	diagnosis	64726704	61053550	94,33%	42338692,5	69,35%
Sample_B2036S	AML	germline	33778194	32086004	94,99%	22891144,5	71,34%
Sample_B2038D	AML	diagnosis	65972137	59428959	90,08%	28658873,5	48,22%
Sample_B2038S	AML	germline	26921438	22798648	84,69%	16305353	71,52%
Sample_B2039D	AML	diagnosis	97163254	90112792	92,74%	61035349,5	67,73%
Sample_B2039S	AML	germline	21081975	19998906	94,86%	14794843	73,98%
Sample_B2040D	AML	diagnosis	64843281	57775681	89,10%	27236564	47,14%
Sample_B2040S	AML	germline	20273825	17307126	85,37%	12479748,5	72,11%
Sample_B2042D	AML	diagnosis	74978250	67526872	90,06%	29543655,5	43,75%
Sample_B2042S	AML	germline	25967438	21220480	81,72%	14405588,5	67,89%
Sample_B2043D	AML	diagnosis	69855875	66050017	94,55%	47102706,5	71,31%
Sample_B2043S	AML	germline	37540729	35563218	94,73%	26371940	74,16%
Sample_B2045D	AML	diagnosis	92623304	86980410	93,91%	58169849,5	66,88%
Sample_B2045S	AML	germline	30542265	24238776	79,36%	16686825	68,84%
Sample_BO_1_NORM	AML	germline	89619176	82202942	91,72%	60051521	73,05%
Sample_BO_1_TUM	AML	diagnosis	98477692	90428863	91,83%	62565696	69,19%
Sample_BO_2_NORM	AML	germline	97607743	89080151	91,26%	76282649,5	85,63%
Sample_BO_2_TUM	AML	diagnosis	75960151	71594138	94,25%	54179288,5	75,68%
Sample_BO_3_NORM	AML	germline	72222284	66890991	92,62%	52344413	78,25%
Sample_BO_3_TUM	AML	diagnosis	109099258	101171451	92,73%	68545995,5	67,75%
Sample_BO_4_NORM	AML	germline	92515480	84090264	90,89%	66530818	79,12%
Sample_BO_4_TUM	AML	diagnosis	78920647	72710447	92,13%	61073425	84,00%
Sample_C0017D	AML	diagnosis	50319837	47517598	94,43%	40631993	85,51%
Sample_C0017S	AML	germline	48194946	45866613	95,17%	37570701,5	81,91%
Sample_C0018D	AML	diagnosis	58984143	55366459	93,87%	49748583,5	89,85%
Sample_C0018S	AML	germline	39221812	37000622	94,34%	33555072	90,69%
Sample_C0022D	AML	diagnosis	35259151	33397766	94,72%	23259070	69,64%

Sample_C0022S	AML	germline	63912405	60340302	94,41%	51156387,5	84,78%
Sample_C0037D	AML	diagnosis	94488008	88275405	93,42%	59374776,5	67,26%
Sample_C0037S	AML	germline	28720805	27268897	94,94%	18412924	67,52%
Sample_C0046D	AML	diagnosis	54449789	49042273	90,07%	23639070,5	48,20%
Sample_C0046S	AML	germline	36549800	28613665	78,29%	20414772	71,35%
Sample_D0027D	AML	diagnosis	97188808	89872158	92,47%	81422047,5	90,60%
Sample_D0027S	AML	germline	27921833	26188675	93,79%	23071968	88,10%
Sample_NGS-41	AML	diagnosis	35044972	33438164	95,42%	30424452	86,82%
Sample_NGS-42	AML	remission	52747850	50362403	95,48%	45318691	85,92%
Sample_NGS-43	AML	diagnosis	62002729	58778197	94,80%	51409926,5	82,92%
Sample_NGS-44	AML	remission	66544386	63623193	95,61%	57161486	85,90%
Sample_NGS-45	AML	diagnosis	60105835	57189266	95,15%	50944690,5	84,76%
Sample_NGS-46	AML	remission	46377420	44352327	95,63%	40157512,5	86,59%
Sample_NGS-47	AML	diagnosis	44710686	42465206	94,98%	38007728,5	85,01%
Sample_NGS-72	AML	remission	56113798	53350952	95,08%	46862905,5	83,51%
Sample_NGS-48	AML	diagnosis	70239377	66932617	95,29%	60072819,5	85,53%
Sample_NGS-50	AML	remission	51133755	48923829	95,68%	43012183,5	84,12%
Sample_NGS-49	AML	diagnosis	45664875	43335085	94,90%	38784926	84,93%
Sample_NGS-51	AML	remission	52551694	50227084	95,58%	44256613	84,22%
Sample_NGS-52	AML	diagnosis	52960850	49843970	94,11%	44606270	84,22%
Sample_NGS-75	AML	remission	49718554	47531547	95,60%	39199407,5	78,84%
Sample_NGS-53	AML	diagnosis	69720610	66279985	95,07%	56391625,5	80,88%
Sample_NGS-58	AML	remission	63391140	60221429	95,00%	52632599,5	83,03%
Sample_NGS-55	AML	diagnosis	51290687	48375549	94,32%	42652115	83,16%
Sample_NGS-62	AML	remission	56043901	52949091	94,48%	46686880	83,30%
Sample_NGS-56	AML	diagnosis	46943999	44496964	94,79%	38372700	81,74%
Sample_NGS-63	AML	remission	54464063	51293141	94,18%	44498496,5	81,70%
Sample_NGS-57	AML	diagnosis	47857451	45511606	95,10%	39548136	82,64%
Sample_NGS-68	AML	remission	46173460	43859468	94,99%	38564247,5	83,52%
Sample_NGS-60	AML	diagnosis	57006627	54066986	94,84%	46294535,5	81,21%
Sample_NGS-64	AML	remission	48716621	46025953	94,48%	40787029	83,72%
Sample_NGS-61	AML	diagnosis	50078379	47566467	94,98%	42000454	83,87%
Sample_NGS-66	AML	remission	46311032	43938076	94,88%	37348548	80,65%
Sample_NGS-65	AML	diagnosis	52686498	49943465	94,79%	44494707,5	84,45%
Sample_NGS-69	AML	remission	57206279	54138645	94,64%	46751609	81,72%
Sample_NGS-67	AML	diagnosis	78211897	73906870	94,50%	61188349	78,23%
Sample_NGS-71	AML	remission	49070392	46753791	95,28%	41374940	84,32%
Sample_NGS-70	AML	diagnosis	54161344	51721877	95,50%	45165111,5	83,39%
Sample_NGS-76	AML	remission	49044273	46051349	93,90%	40146693	81,86%
Sample_NGS-73	AML	diagnosis	46723571	44456325	95,15%	39157963,5	83,81%
Sample_NGS-78	AML	remission	51779282	49243850	95,10%	42487234,5	82,05%
Sample_NGS-74	AML	diagnosis	43731754	41819716	95,63%	36505232,5	83,48%
Sample_NGS-77	AML	remission	58431356	55362472	94,75%	48568769,5	83,12%
Sample_NGS-79	AML	diagnosis	52543744	49838707	94,85%	43896050	83,54%
Sample_NGS-86	AML	remission	58501849	55233848	94,41%	47461781,5	81,13%
Sample_NGS-80	AML	diagnosis	54283878	51514772	94,90%	44637854,5	82,23%
Sample_NGS-84	AML	remission	47408554	44922686	94,76%	37966851,5	80,08%

Sample_NGS-81	AML	diagnosis	47142236	44761008	94,95%	38091263	80,80%
Sample_NGS-83	AML	remission	61827641	58288808	94,28%	49732448	80,44%
Sample_NGS-82	AML	diagnosis	53496665	50819253	95,00%	43817174	81,91%
Sample_NGS-87	AML	remission	57795915	54589073	94,45%	47068972	81,44%
Sample_3FK_3D	ALL	diagnosis	69102695	63510395	91,91%	56638083,5	81,96%
Sample_3FK_3n-DNA	ALL	germline	71168290	68044617	95,61%	44299504,5	65,10%
Sample_3FK_3R	ALL	relapse	74056049	68897070	93,03%	60441971,5	81,62%
Sample_4PJ_4D	ALL	diagnosis	82005678	75615433	92,21%	66561836,5	81,17%
Sample_4PJ_4n-DNA	ALL	germline	83522059	78570586	94,07%	67061889	85,35%
Sample_4PJ_4R	ALL	relapse	56120187	52157901	92,94%	47677698,5	84,96%
Sample_6MJ_6D	ALL	diagnosis	93796558	88726039	94,59%	63424179	67,62%
Sample_6MJ_6n-DNA	ALL	germline	74594066	70517430	94,53%	55372132,5	74,23%
Sample_7TK_7D	ALL	diagnosis	115546871	110698372	95,80%	87264041	75,52%
Sample_7TK_7n-DNA	ALL	germline	31508710	30317788	96,22%	22355056,5	70,95%
Sample_8PB_8D	ALL	diagnosis	144493912	138250366	95,68%	103439036,5	71,59%
Sample_8PB_8n-DNA	ALL	germline	33950946	32565043	95,92%	24250220	71,43%
Sample_10JN_10D	ALL	diagnosis	78282405	73476504	93,86%	63715601	81,39%
Sample_10JN_10n-DNA	ALL	germline	52395970	50581368	96,54%	41205248	78,64%
Sample_10JN_10R	ALL	relapse	52179625	50251503	96,30%	41490644,5	79,52%
Sample_11LT_11D	ALL	diagnosis	68171713	63343797	92,92%	56070575,5	82,25%
Sample_11LT_11n-DNA	ALL	germline	107990339	102501545	94,92%	84247953,5	82,19%
Sample_554	ALL	diagnosis	76073023	70944161	93,26%	59082843,5	77,67%
Sample_1629	ALL	remission	66743298	62140898	93,10%	52466472	78,61%
Sample_616	ALL	diagnosis	65266981	62255759	95,39%	53505595	81,98%
Sample_1630	ALL	remission	86137139	80646257	93,63%	67370410	78,21%
Sample_757	ALL	diagnosis	58174042	54864088	94,31%	46323110,5	79,63%
Sample_751	ALL	germline	79226798	74952067	94,60%	61823282	78,03%
Sample_961	ALL	relapse	67814593	64151051	94,60%	53359211,5	78,68%
Sample_1009	ALL	germline	64266221	59803150	93,06%	50539030,5	78,64%
Sample_960	ALL	diagnosis	50432072	48064597	95,31%	41167518,5	81,63%
Sample_1011	ALL	germline	43915140	40851423	93,02%	34663238,5	78,93%
Sample_1258	ALL	diagnosis	37795346	35867488	94,90%	31457757,5	83,23%
Sample_1341	ALL	germline	47589113	44131119	92,73%	36882800	77,50%
Sample_1430	ALL	diagnosis	69212642	65038346	93,97%	57472172	83,04%
Sample_1612	ALL	germline	39806959	36424628	91,50%	30022563	75,42%
Sample_1731	ALL	diagnosis	78217014	73709543	94,24%	40469026,5	54,90%
Sample_1764	ALL	germline	71775242	67464261	93,99%	40067253,5	59,39%
Sample_30846	ALL	diagnosis	22842232	21120132	92,46%	16758299,5	73,37%
Sample_37839	ALL	remission	73833615	69290082	93,85%	59096982,5	80,04%
Sample_43873	ALL	diagnosis	47264247	44218792	93,56%	38736814,5	81,96%
Sample_44365	ALL	remission	55070575	51881778	94,21%	44949302	81,62%
Sample_65420	ALL	diagnosis	61844155	58419603	94,46%	50899102,5	82,30%
Sample_80535	ALL	remission	33797942	32296904	95,56%	28963946,5	85,70%
Sample_74413	ALL	diagnosis	61456436	56896952	92,58%	50473434,5	82,13%

Sample_75147	ALL	remission	66498953	62801775	94,44%	54186091,5	81,48%
Sample_78540	ALL	diagnosis	70630741	66556740	94,23%	56945706,5	80,62%
Sample_79323	ALL	remission	61267898	57987996	94,65%	50937543	83,14%
Sample_85112_8511	ALL	diagnosis	56200834	52238366	92,95%	46670675,5	83,04%
Sample_295012_2950	ALL	remission	84045875	79493727	94,58%	61526410	77,40%
Sample_106013_1060	ALL	diagnosis	72166868	67908483	94,10%	57157813,5	79,20%
Sample_125613_1256	ALL	remission	78404359	74799886	95,40%	62075663	79,17%
Sample_108612_1086	ALL	diagnosis	61520145	56346898	91,59%	41570830,5	67,57%
Sample_163213_1632	ALL	remission	73814210	70038317	94,88%	58797903,5	83,95%
Sample_139213_1392	ALL	diagnosis	72104183	68514893	95,02%	57841350,5	84,42%
Sample_206613_2066	ALL	remission	82033667	74887273	91,29%	63856239	77,84%
Sample_246313_2463	ALL	diagnosis	65220624	59933071	91,89%	45285724	69,43%
Sample_222313_2223	ALL	remission	72936583	68546854	93,98%	57605101	84,04%
Sample_331212_3312	ALL	diagnosis	83285708	78051144	93,71%	65988252,5	84,54%
Sample_9813_98	ALL	remission	82262913	75133585	91,33%	63471828,5	77,16%
Sample_417612_4176	ALL	diagnosis	99577757	90801235	91,19%	76855088,5	77,18%
Sample_220313_2203	ALL	remission	73721238	69197354	93,86%	58648002,5	84,75%
Sample_NGS-163	ALL	diagnosis	61600192	57851857	93,92%	50864258	82,57%
Sample_NGS-164	ALL	remission	75096097	71175145	94,78%	61436921,5	81,81%
Sample_NGS-165	ALL	diagnosis	76933888	72402718	94,11%	61987837	80,57%
Sample_NGS-166	ALL	remission	84328301	79798479	94,63%	68830342	81,62%
Sample_NGS-167	ALL	diagnosis	75085497	70816928	94,32%	61954525	82,51%
Sample_NGS-168	ALL	remission	73798144	69516131	94,20%	60531899,5	82,02%
Sample_NGS-169	ALL	diagnosis	80236416	75162117	93,68%	66265720	82,59%
Sample_NGS-170	ALL	remission	81078668	76072642	93,83%	65590191,5	80,90%
Sample_NGS-171	ALL	diagnosis	82945032	77645159	93,61%	67768557	81,70%
Sample_NGS-172	ALL	remission	77979291	73287041	93,98%	62700968	80,41%
Sample_NGS-173	ALL	diagnosis	69375244	64936228	93,60%	54730042	78,89%
Sample_NGS-174	ALL	remission	71649494	67329539	93,97%	57425696,5	80,15%
Sample_NGS-175	ALL	diagnosis	74944867	69980509	93,38%	61626636,5	82,23%
Sample_NGS-176	ALL	remission	84447221	78745042	93,25%	68133005,5	80,68%
Sample_NGS-177	ALL	diagnosis	75834804	70670030	93,19%	60997932	80,44%
Sample_NGS-178	ALL	remission	98919493	92024101	93,03%	81883933	82,78%
Sample_NGS-179	ALL	diagnosis	72307302	67221669	92,97%	58799351,5	81,32%
Sample_NGS-180	ALL	remission	64732993	60461248	93,40%	53127906	82,07%
Sample_NGS-183	ALL	diagnosis	85964106	79844085	92,88%	69671995,5	87,26%
Sample_NGS-185	ALL	germline	74741371	69352722	92,79%	60146480	86,73%
		MEAN	61672249,64	57795861,64	93,64%	47478769,62	80,01%

APPENDIX 2

Detailed description on average coverage after filtering and deduplication of the fragments and the percentage of target bases covered by at least 1, 10, 20 reads.

SAMPLE	Type of Leukemia	Phase	Coverage Mean	Covered 1x	Covered 10x	Covered 20x
Sample_187	AML	diagnosis	94,91	95,05%	87,70%	81,82%
Sample_197	AML	germline	83,6	95,09%	87,67%	81,59%
Sample_195	AML	diagnosis	84,39	94,70%	86,22%	78,83%
Sample_198	AML	germline	95,49	95,61%	88,76%	83,47%
Sample_63640	AML	diagnosis	98,95	95,32%	88,59%	83,61%
Sample_199	AML	germline	37,64	92,67%	76,16%	59,54%
Sample_A1010D	AML	diagnosis	94,35	94,31%	87,64%	83,37%
Sample_A1010S	AML	germline	120,94	95,00%	89,03%	85,71%
Sample_A1015Dbis	AML	diagnosis	73,51	97,11%	93,08%	89,12%
Sample_A1015S	AML	germline	33,65	96,10%	85,55%	65,10%
Sample_A1024D	AML	diagnosis	112,66	97,76%	94,46%	91,94%
Sample_A1024S	AML	germline	42,39	96,54%	88,14%	72,30%
Sample_A1025D	AML	diagnosis	112,47	97,66%	94,41%	91,92%
Sample_A1025S	AML	germline	49,97	96,98%	90,61%	79,26%
Sample_B1001D	AML	diagnosis	133,16	95,44%	89,68%	86,55%
Sample_B1001S	AML	germline	95,53	94,75%	88,21%	84,18%
Sample_B1006D	AML	diagnosis	82,19	97,15%	93,48%	89,94%
Sample_B1006S	AML	germline	69,05	97,50%	93,03%	87,91%
Sample_B1014D	AML	diagnosis	72,94	97,54%	93,17%	89,34%
Sample_B1014S	AML	germline	34,14	96,46%	87,99%	72,72%
Sample_B1026D	AML	diagnosis	74,57	97,14%	93,10%	88,99%
Sample_B1026S	AML	germline	45,15	96,47%	89,37%	76,53%
Sample_B1028D	AML	diagnosis	98,69	97,21%	93,77%	90,75%
Sample_B1028S	AML	germline	44,54	96,41%	89,52%	77,16%
Sample_B1034D	AML	diagnosis	67,45	96,88%	92,78%	88,19%
Sample_B1034S	AML	germline	48,91	96,59%	89,88%	78,20%
Sample_B1041D	AML	diagnosis	78,94	97,17%	89,52%	81,58%
Sample_B1041S	AML	germline	44,51	94,57%	76,91%	61,52%
Sample_B2002D	AML	diagnosis	82,46	97,25%	93,48%	89,78%
Sample_B2002S	AML	germline	29,89	95,99%	82,64%	58,92%
Sample_B2004D	AML	diagnosis	81,72	96,04%	90,67%	85,98%
Sample_B2004S	AML	germline	42,1	96,69%	89,94%	79,30%
Sample_B2005D	AML	diagnosis	83,34	97,18%	93,05%	89,24%
Sample_B2005S	AML	germline	85,2	97,66%	93,47%	89,84%
Sample_B2007D	AML	diagnosis	95,64	97,56%	94,19%	91,10%
Sample_B2007S	AML	germline	70,33	96,14%	90,30%	85,16%
Sample_B2008D	AML	diagnosis	145,33	94,19%	88,58%	85,60%
Sample_B2008S	AML	germline	98,77	94,23%	87,58%	83,45%
Sample_B2009D	AML	diagnosis	85,45	97,27%	93,29%	89,30%
Sample_B2009S	AML	germline	36,81	96,74%	88,77%	74,72%

Sample_B2023D	AML	diagnosis	73,05	97,17%	93,13%	89,08%
Sample_B2023S	AML	germline	55,79	96,78%	90,99%	81,51%
Sample_B2030D	AML	diagnosis	111,22	97,33%	94,11%	91,47%
Sample_B2030S	AML	germline	44,95	96,52%	89,04%	75,07%
Sample_B2031D	AML	diagnosis	58,48	96,84%	92,16%	86,85%
Sample_B2031S	AML	germline	45,93	96,78%	90,01%	78,23%
Sample_B2033D	AML	diagnosis	81,08	97,15%	93,41%	89,73%
Sample_B2033S	AML	germline	42,91	96,44%	88,58%	73,98%
Sample_B2035D	AML	diagnosis	50,41	96,69%	85,36%	69,42%
Sample_B2035S	AML	germline	29,93	94,41%	69,95%	47,53%
Sample_B2036D	AML	diagnosis	75,14	97,37%	89,70%	81,55%
Sample_B2036S	AML	germline	47,5	94,56%	77,85%	63,01%
Sample_B2038D	AML	diagnosis	55,5	96,80%	87,09%	73,14%
Sample_B2038S	AML	germline	32,88	95,35%	76,24%	53,29%
Sample_B2039D	AML	diagnosis	108,58	97,96%	92,24%	87,44%
Sample_B2039S	AML	germline	30,64	94,48%	72,38%	52,39%
Sample_B2040D	AML	diagnosis	52,88	96,89%	86,57%	71,24%
Sample_B2040S	AML	germline	25,18	94,56%	68,19%	42,91%
Sample_B2042D	AML	diagnosis	58,38	96,75%	86,86%	72,82%
Sample_B2042S	AML	germline	29,33	95,08%	71,46%	47,59%
Sample_B2043D	AML	diagnosis	81,73	97,51%	90,48%	83,44%
Sample_B2043S	AML	germline	52,59	96,22%	84,86%	70,47%
Sample_B2045D	AML	diagnosis	101,18	97,64%	91,29%	85,91%
Sample_B2045S	AML	germline	34,05	95,61%	77,40%	54,06%
Sample_BO_1_NORM	AML	germline	165,58	99,62%	98,75%	96,97%
Sample_BO_1_TUM	AML	diagnosis	172,8	99,59%	98,60%	96,71%
Sample_BO_2_NORM	AML	germline	218,14	99,61%	99,12%	98,46%
Sample_BO_2_TUM	AML	diagnosis	147,06	99,59%	98,44%	96,08%
Sample_BO_3_NORM	AML	germline	143,78	99,52%	98,42%	96,16%
Sample_BO_3_TUM	AML	diagnosis	187,36	99,52%	98,69%	97,14%
Sample_BO_4_NORM	AML	germline	140,32	99,50%	98,32%	95,85%
Sample_BO_4_TUM	AML	diagnosis	166,12	99,52%	98,47%	96,52%
Sample_C0017D	AML	diagnosis	68,72	93,59%	85,82%	80,12%
Sample_C0017S	AML	germline	67,59	93,66%	85,96%	80,67%
Sample_C0018D	AML	diagnosis	81,1	97,39%	93,39%	89,83%
Sample_C0018S	AML	germline	56,6	96,71%	91,40%	83,29%
Sample_C0022D	AML	diagnosis	39,66	92,92%	82,71%	73,31%
Sample_C0022S	AML	germline	79,37	94,21%	86,59%	81,57%
Sample_C0037D	AML	diagnosis	102,52	97,88%	91,88%	87,09%
Sample_C0037S	AML	germline	38,51	92,91%	70,39%	55,72%
Sample_C0046D	AML	diagnosis	46,33	96,24%	82,91%	64,54%
Sample_C0046S	AML	germline	40,75	96,54%	82,43%	62,12%
Sample_D0027D	AML	diagnosis	133,84	97,77%	94,47%	92,36%
Sample_D0027S	AML	germline	38,46	96,27%	88,22%	74,06%
Sample_NGS-41	AML	diagnosis	51.96	96,19%	88,81%	78,60%
Sample_NGS-42	AML	remission	80.28	96,60%	91,49%	86,87%
Sample_NGS-43	AML	diagnosis	87.66	97,05%	92,19%	88,08%

Sample_NGS-44	AML	remission	99.82	96,94%	92,48%	88,90%
Sample_NGS-45	AML	diagnosis	87.99	96,85%	92,01%	87,86%
Sample_NGS-46	AML	remission	67.97	96,63%	90,79%	84,77%
Sample_NGS-47	AML	diagnosis	64.64	96,90%	90,98%	84,53%
Sample_NGS-72	AML	remission	73.79	97,28%	91,77%	86,54%
Sample_NGS-48	AML	diagnosis	93.29	97,71%	92,50%	88,52%
Sample_NGS-50	AML	remission	73.55	96,79%	91,46%	86,19%
Sample_NGS-49	AML	diagnosis	61.47	96,96%	90,96%	84,35%
Sample_NGS-51	AML	remission	73.72	96,83%	91,53%	86,53%
Sample_NGS-52	AML	diagnosis	76.74	96,86%	91,56%	86,58%
Sample_NGS-75	AML	remission	65.45	96,85%	91,06%	85,35%
Sample_NGS-53	AML	diagnosis	97.43	96,92%	92,52%	89,19%
Sample_NGS-58	AML	remission	89.76	96,95%	92,20%	88,49%
Sample_NGS-55	AML	diagnosis	75.76	96,58%	91,36%	86,57%
Sample_NGS-62	AML	remission	80.99	96,80%	91,77%	87,39%
Sample_NGS-56	AML	diagnosis	67.36	96,46%	90,75%	84,63%
Sample_NGS-63	AML	remission	77.82	96,47%	91,30%	86,55%
Sample_NGS-57	AML	diagnosis	68.47	96,87%	91,31%	85,52%
Sample_NGS-68	AML	remission	66.97	96,62%	91,14%	85,38%
Sample_NGS-60	AML	diagnosis	81.18	96,63%	91,60%	86,91%
Sample_NGS-64	AML	remission	71.2	96,49%	91,29%	86,09%
Sample_NGS-61	AML	diagnosis	72.1	96,71%	91,40%	86,10%
Sample_NGS-66	AML	remission	65.31	96,65%	90,81%	84,87%
Sample_NGS-65	AML	diagnosis	76.65	96,71%	91,35%	86,25%
Sample_NGS-69	AML	remission	81.39	96,72%	91,61%	86,90%
Sample_NGS-67	AML	diagnosis	104.59	97,11%	92,88%	89,79%
Sample_NGS-71	AML	remission	71.57	96,75%	91,23%	85,62%
Sample_NGS-70	AML	diagnosis	75.87	96,84%	91,59%	86,74%
Sample_NGS-76	AML	remission	64.91	97,10%	91,70%	85,74%
Sample_NGS-73	AML	diagnosis	62.81	96,98%	90,78%	83,69%
Sample_NGS-78	AML	remission	72.45	96,93%	91,43%	85,96%
Sample_NGS-74	AML	diagnosis	60.37	96,89%	90,72%	83,30%
Sample_NGS-77	AML	remission	82.65	97,08%	91,86%	86,93%
Sample_NGS-79	AML	diagnosis	76.59	96,78%	91,55%	86,55%
Sample_NGS-86	AML	remission	82.14	96,99%	92,07%	87,56%
Sample_NGS-80	AML	diagnosis	77.79	96,84%	91,56%	86,72%
Sample_NGS-84	AML	remission	65.29	96,82%	91,05%	84,95%
Sample_NGS-81	AML	diagnosis	65.52	96,76%	90,91%	84,75%
Sample_NGS-83	AML	remission	83.64	97,18%	92,57%	88,48%
Sample_NGS-82	AML	diagnosis	74.9	96,84%	91,59%	86,66%
Sample_NGS-87	AML	remission	80.89	96,85%	91,88%	87,51%
Sample_3FK_3D	ALL	diagnosis	97,23	97,28%	92,88%	89,44%
Sample_3FK_3n-DNA	ALL	germline	79,52	96,56%	91,54%	84,87%
Sample_3FK_3R	ALL	relapse	104,01	97,14%	92,98%	89,58%
Sample_4PJ_4D	ALL	diagnosis	106,45	97,76%	93,18%	90,15%
Sample_4PJ_4n-DNA	ALL	germline	118,41	97,12%	93,28%	90,37%
Sample_4PJ_4R	ALL	relapse	76,45	97,34%	92,77%	87,51%

Sample_6MJ_6D	ALL	diagnosis	111,26	97,09%	93,60%	91,13%
Sample_6MJ_6n-DNA	ALL	germline	97,89	97,11%	93,07%	89,82%
Sample_7TK_7D	ALL	diagnosis	148,81	97,74%	94,68%	92,57%
Sample_7TK_7n-DNA	ALL	germline	40,71	95,68%	87,20%	72,64%
Sample_8PB_8D	ALL	diagnosis	171,94	98,16%	95,16%	93,28%
Sample_8PB_8n-DNA	ALL	germline	44,61	95,35%	82,91%	66,91%
Sample_10JN_10D	ALL	diagnosis	106,68	97,40%	92,79%	89,52%
Sample_10JN_10n-DNA	ALL	germline	75,12	96,46%	91,37%	86,35%
Sample_10JN_10R	ALL	relapse	74,53	96,68%	91,66%	86,73%
Sample_11LT_11D	ALL	diagnosis	96,88	96,86%	92,47%	88,93%
Sample_11LT_11n-DNA	ALL	germline	147,17	97,14%	92,38%	87,33%
Sample_554	ALL	diagnosis	101,94	97,06%	92,67%	89,54%
Sample_1629	ALL	remission	93,51	96,88%	92,23%	88,59%
Sample_616	ALL	diagnosis	92,17	96,98%	92,27%	88,59%
Sample_1630	ALL	remission	118,89	97,24%	93,14%	90,40%
Sample_757	ALL	diagnosis	80,27	96,77%	91,68%	87,24%
Sample_751	ALL	germline	113,37	96,62%	91,88%	88,69%
Sample_961	ALL	relapse	91,90	97,11%	92,42%	88,72%
Sample_1009	ALL	germline	90,52	96,63%	91,82%	87,90%
Sample_960	ALL	diagnosis	71,34	96,56%	91,20%	85,99%
Sample_1011	ALL	germline	60,94	96,25%	89,94%	82,41%
Sample_1258	ALL	diagnosis	54,47	96,07%	89,67%	81,60%
Sample_1341	ALL	germline	64,50	96,52%	90,67%	84,43%
Sample_1430	ALL	diagnosis	100,66	97,00%	92,43%	88,85%
Sample_1612	ALL	germline	52,95	96,20%	89,98%	82,35%
Sample_1731	ALL	diagnosis	80,81	95,02%	86,82%	79,29%
Sample_1764	ALL	germline	80,27	95,28%	87,57%	80,39%
Sample_30846	ALL	diagnosis	28,32	96,00%	78,22%	52,40%
Sample_37839	ALL	remission	105,28	96,87%	92,53%	89,43%
Sample_43873	ALL	diagnosis	62,19	97,32%	90,72%	82,83%
Sample_44365	ALL	remission	75,22	97,05%	91,63%	86,52%
Sample_65420	ALL	diagnosis	89,17	96,76%	91,92%	87,91%
Sample_80535	ALL	remission	49,95	96,27%	89,17%	79,30%
Sample_74413	ALL	diagnosis	89,47	96,71%	92,01%	88,14%
Sample_75147	ALL	remission	96,55	96,85%	92,29%	88,86%
Sample_78540	ALL	diagnosis	99,56	97,12%	92,50%	88,90%
Sample_79323	ALL	remission	85,59	97,24%	92,07%	87,54%
Sample_85112_8511	ALL	diagnosis	80,83	96,81%	91,97%	87,70%
Sample_295012_2950	ALL	remission	108,78	97,13%	93,28%	90,47%
Sample_106013_1060	ALL	diagnosis	96,33	97,12%	92,95%	89,31%
Sample_125613_1256	ALL	remission	107,96	97,03%	93,19%	90,19%
Sample_108612_1086	ALL	diagnosis	71,70	96,88%	92,07%	87,69%
Sample_163213_1632	ALL	remission	102,49	97,04%	93,11%	89,92%
Sample_139213_1392	ALL	diagnosis	98,17	97,21%	93,11%	89,49%
Sample_206613_2066	ALL	remission	108,11	97,34%	93,48%	90,56%
Sample_246313_2463	ALL	diagnosis	78,73	96,87%	92,03%	87,67%

Sample_222313_2223	ALL	remission	98,54	97,19%	93,26%	89,91%
Sample_331212_3312	ALL	diagnosis	112,34	97,26%	93,42%	90,41%
Sample_9813_98	ALL	remission	109,34	97,24%	93,29%	90,51%
Sample_417612_4176	ALL	diagnosis	133,53	97,39%	93,66%	91,36%
Sample_220313_2203	ALL	remission	100,09	97,17%	93,27%	90,07%
Sample_NGS-163	ALL	diagnosis	86,93	96,97%	92,66%	88,61%
Sample_NGS-164	ALL	remission	105,63	97,12%	93,09%	89,98%
Sample_NGS-165	ALL	diagnosis	105,30	97,12%	93,16%	89,88%
Sample_NGS-166	ALL	remission	118,13	97,21%	93,49%	90,70%
Sample_NGS-167	ALL	diagnosis	106,96	96,96%	92,98%	89,89%
Sample_NGS-168	ALL	remission	104,68	96,96%	93,01%	89,98%
Sample_NGS-169	ALL	diagnosis	113,20	97,20%	93,23%	90,28%
Sample_NGS-170	ALL	remission	113,99	97,01%	93,08%	90,07%
Sample_NGS-171	ALL	diagnosis	114,53	97,21%	93,39%	90,43%
Sample_NGS-172	ALL	remission	107,10	97,25%	93,40%	90,39%
Sample_NGS-173	ALL	diagnosis	91,97	97,14%	92,91%	89,22%
Sample_NGS-174	ALL	remission	99,24	97,12%	93,11%	89,80%
Sample_NGS-175	ALL	diagnosis	105,56	97,20%	93,20%	89,99%
Sample_NGS-176	ALL	remission	114,65	97,42%	93,52%	90,65%
Sample_NGS-177	ALL	diagnosis	105,37	97,09%	93,11%	90,09%
Sample_NGS-178	ALL	remission	125,59	97,97%	93,67%	91,10%
Sample_NGS-179	ALL	diagnosis	92,72	97,55%	92,81%	89,15%
Sample_NGS-180	ALL	remission	89,02	97,36%	92,78%	88,77%
Sample_NGS-183	ALL	diagnosis	118,25	97,40%	93,63%	90,80%
Sample_NGS-185	ALL	germline	105,03	97,02%	93,07%	89,94%
		MEAN	86,54608696	96,73%	90,55%	84,09%

REFERENCES

[1] D. Hanahan and R. A. Weinberg, ‘Hallmarks of cancer: The next generation’, *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[2] M. R. Stratton, P. J. Campbell, and P. A. Futreal, ‘The cancer genome.’, *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[3] M. Greaves and C. C. Maley, ‘Clonal evolution in cancer’, *Nature*, vol. 481, no. 7381, pp. 306–313, 2012.

[4] M. R. Stratton, ‘Journeys into the genome of cancer cells’, *EMBO Mol. Med.*, vol. 5, no. 2, pp. 169–172, 2013.

[5] G. a Koretzky, ‘Review series introduction The legacy of the Philadelphia chromosome’, *J. Clin. Invest.*, vol. 117, no. 8, pp. 2030–2032, 2007.

[6] P. Baliakas *et al.*, ‘Recurrent mutations refine prognosis in chronic lymphocytic leukemia.’, *Leukemia*, no. April, pp. 1–8, 2014.

[7] A. Puiggros, G. Blanco, and B. Espinet, ‘Genetic abnormalities in chronic lymphocytic leukemia: Where we are and where we go’, *Biomed Res. Int.*, vol. 2014, 2014.

[8] T. Naoe and H. Kiyoi, ‘Gene mutations of acute myeloid leukemia in the genome era’, *Int. J. Hematol.*, vol. 97, no. 2, pp. 165–174, 2013.

[9] A. H. Shih, O. Abdel-Wahab, J. P. Patel, and R. L. Levine, ‘The role of mutations in epigenetic regulators in myeloid malignancies’, *Nat Rev Cancer*, vol. 12, no. 9, pp. 599–612, 2012.

[10] C. G. Mullighan, ‘The molecular genetic makeup of acute lymphoblastic leukemia.’, *Hematology Am. Soc. Hematol. Educ. Program*, vol. 2012, pp. 389–96, 2012.

[11] M. Choi *et al.*, ‘Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.’, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 45, pp. 19096–101, 2009.

[12] D. C. Koboldt *et al.*, ‘VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing’, *Genome Res.*, vol. 22, no. 3, pp. 568–576, 2012.

[13] C. T. Saunders, W. S. W. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham, ‘Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs’, *Bioinformatics*, vol. 28, no. 14, pp. 1811–1817, 2012.

[14] K. Cibulskis *et al.*, ‘Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples’, *Nat. Biotechnol.*, vol. 31, no. 3, pp. 213–219, 2013.

[15] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, ‘SIFT web server: predicting effects of amino acid substitutions on proteins.’, *Nucleic Acids Res.*, vol. 40, no. Web Server issue, pp. W452-7, 2012.

[16] I. A. Adzhubei *et al.*, ‘A method and server for predicting damaging missense mutations’, *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[17] E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, and L. Cai, ‘MutationTaster2: mutation prediction for the deep-sequencing age’, *Nat. Methods Nat. Methods Nat. Methods*, vol. 9, no. 10, pp. 743–748, 2012.

[18] B. Reva, Y. Antipin, and C. Sander, ‘Predicting the functional impact of protein mutations: Application to cancer genomics’, *Nucleic Acids Res.*, vol. 39, no. 17, pp. 37–43, 2011.

[19] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, ‘Identifying a high fraction of the human genome to be under selective constraint using GERP++’, *PLoS Comput. Biol.*, vol. 6, no. 12, 2010.

[20] M. Lek *et al.*, ‘Analysis of protein-coding genetic variation in 60,706 humans’, *Nature*, vol. 536, no. 7616, pp. 285–291, 2016.

[21] A. Auton *et al.*, ‘A global reference for human genetic variation’, *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[22] W. Fu *et al.*, ‘Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants’, *Nature*, vol. 493, no. 7431, pp. 216–220, 2012.

[23] S. Bamford *et al.*, ‘The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website.’, *Br. J. Cancer*, vol. 91, no. 2, pp. 355–8, 2004.

[24] S. A. Forbes *et al.*, ‘COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer’, *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, pp. 945–950, 2011.

[25] J. Zhang *et al.*, ‘International cancer genome consortium data portal-a one-stop shop for cancer genomics data’, *Database*, vol. 2011, pp. 1–10, 2011.

[26] C. A. Ramirez *et al.*, ‘CIViC: A knowledgebase for expert-crowdsourcing the clinical interpretation of variants in cancer’, pp. 4–11, 2016.

[27] M. S. Lawrence *et al.*, ‘Mutational heterogeneity in cancer and the search for new cancer-associated genes.’, *Nature*, vol. 499, pp. 214–8, 2013.

[28] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas, ‘OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes’, *Bioinformatics*, vol. 29, no. 18, pp. 2238–2244, 2013.

[29] A. Gonzalez-Perez and N. Lopez-Bigas, ‘Functional impact bias reveals cancer drivers’, *Nucleic Acids Res.*, vol. 40, no. 21, pp. 1–10, 2012.

[30] C. M. Zwaan *et al.*, ‘Collaborative efforts driving progress in pediatric acute myeloid leukemia’, *J. Clin. Oncol.*, vol. 33, no. 27, pp. 2949–2962, 2015.

[31] M. A. DePristo *et al.*, ‘A framework for variation discovery and genotyping using next-generation DNA sequencing data.’, *Nat. Genet.*, vol. 43, no. 5, pp. 491–8, 2011.

[32] R. K. Patel and M. Jain, ‘NGS QC toolkit: A toolkit for quality control of next generation sequencing data’, *PLoS One*, vol. 7, no. 2, 2012.

[33] M. D’Antonio *et al.*, ‘WEP: a high-performance analysis pipeline for whole-exome data’, *BMC Bioinformatics*, vol. 14, no. Suppl 7, p. S11, 2013.

[34] H. Li and R. Durbin, ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[35] G. A. Van der Auwera *et al.*, *From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*, no. SUPL.43. 2013.

[36] S. T. Sherry *et al.*, ‘dbSNP: the NCBI database of genetic variation.’, *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, 2001.

[37] N. A. O’Leary *et al.*, ‘Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation’, *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, 2016.

[38] X. Liu, C. Wu, C. Li, and E. Boerwinkle, ‘dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs’, *Hum. Mutat.*, vol. 37, no. 3, pp. 235–241, 2016.

[39] K. Wang, M. Li, and H. Hakonarson, ‘ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data’, *Nucleic Acids Res.*, vol. 38, no. 16, p. e164, 2010.

[40] L. Qiu *et al.*, ‘Downregulation of CDC27 inhibits the proliferation of colorectal cancer cells via the accumulation of p21Cip1/Waf1’, *Cell Death Dis.*, vol. 7, no. 1, p. e2074, 2016.

[41] L. Ding *et al.*, ‘Somatic mutations affect key pathways in lung

adenocarcinoma.', *Nature*, vol. 455, no. 7216, pp. 1069–75, 2008.

[42] S. I. Nikolaev *et al.*, 'Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma', *Nat. Genet.*, vol. 44, no. 2, pp. 133–139, 2012.

[43] D. W. Craig *et al.*, 'Genome and Transcriptome Sequencing in Prospective Metastatic Triple-Negative Breast Cancer Uncovers Therapeutic Vulnerabilities', *Mol. Cancer Ther.*, vol. 12, no. 1, pp. 104–116, 2013.

[44] H. Now and J. Y. Yoo, 'A protein-kinase, IFN-inducible double-stranded RNA dependent inhibitor and repressor of p58 (PRKRIR) enhances type I IFN-mediated antiviral response through the stability control of RIG-I protein', *Biochem. Biophys. Res. Commun.*, vol. 413, no. 3, pp. 487–493, 2011.