PhD degree in Molecular Medicine
Curriculum in Molecular Oncology
European School of Molecular Medicine (SEMM)
University of Milan and University of Naples "Federico II"
Faculty of Medicine
Settore disciplinare MED/4

Clonal tracking and high throughput shRNA screening in AMLs

Umberto Andrea Cammarata
IFOM-IEO Campus, Milan
Registration Number n. R09410

Supervisor: **Prof. Pier Giuseppe Pelicci**
Dr. Emanuela Colombo
Dr. Chiara Ronchini

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

5FU	5-Fluorouracil
Alb	Albumin
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
APL	Acute Promyelocytic Leukemia
ATRA	Trans-Retinoic Acid
BM	Bone Marrow
CAN	Copy Number Alteration
CLL	Chronic Lymphocytic Leukemia
CLP	Common Lymphoid Progenitor
CML	Chronic Myeloid Leukemia
CMP	Common Myeloid Progenitors
CR	Complete Remission
CSC	Cancer Stem Cell
Ct	Cycle Threshold
DMSO	Dimethyl Sulfoxide
Dox	Doxycycline
ELDA	Extreme Limiting Dilution Analysis
FA	Formaldehyde
FAB	French-American-British
FACS	Fluorescence-Activated Cell Sorting
FBS	Foetal Bovine Serum
FC	Fold Change
FISH	Fluorescence In Situ Hybridization
GFP	Green Fluorescent Protein
GMP	Granulocytes-Macrophage Progenitors

HSC	Hematopoietic Stem Cell
HTS	High Throughput Screening
KI	Knock In
LB	Luria Broth
LDT	Limiting Dilution Transplantation
LIC	Leukemia Initiating Cell
LT-HSC	Long Term-Hematopoietic Stem Cell
MEP	Megakaryocyte-Erythrocytes Progenitors
miRNA	MicroRNA
MOI	Multiplicity of Infection
MPP	Multipotent Progenitor
NGS	Next Generation Sequencing
NK	Natural Killer
NSG	Nod.Cg-Prkdcscid Il2rgtm1wjl/Szj
O/N	Over Night
OS	Overall Survival Rate
PB	Peripheral Blood
PML	Promyelocytic Protein
PTK	Protein Tyrosine Kinases
qPCR	Quantitative PCR
RA	Retinoic Acid
RARa	Retinoic Acid Receptor-Alpha
RBC	Red Blood Cell
RFP	Red Fluorescent Protein
RISC	RNA-Induced Silencing Complex
RNAi	RNA Interference
RT	Room Temperature

shRNA	Short-Hairpin RNA
siRNA	Interfering RNA
SPL	Spleen
ST-HSC	Short Term-Hematopoietic Stem Cell
TAE	Tris-Acetate-EDTA
TB	Terrific Broth
TRES2	Tetracycline Response Element 2
TU	Transducing Units
WGS	Whole Genome Sequencing
WHO	World Health Organization

ABSTRACT

Acute myeloid leukemia (AML) is a malignant tumor characterized by the uncontrolled proliferation of immature hematopoietic cells (blasts) that colonize the bone marrow leading to a malfunctional hematopoiesis. Despite the efforts to search for new therapies and the identification of new targets, leukemias are still largely incurable. The main reason for treatment failure is the high frequency of relapses after remission induced by the first chemotherapy treatment. Recent studies suggest that the high rate of relapse is due to the presence of cells with limitless regenerative capacity, Leukemia Initiating Cells (LIC), that are very resistant to chemotherapy. Moreover, since the leukemic blasts are highly heterogeneous both from a biological and genetic point of view, AML evolves continuously under both environmental and external stimuli (eg, chemotherapy).

In order to study the clonal evolution of AML and to find new therapeutic targets that allow to eradicate leukemic blasts, this study has set the following objectives: i) monitor LICs growth *in vivo* through the transplantation of traceable leukemic cells; ii) investigate the clonal evolution of the disease; iii) identify genes critical to tumor growth *in vivo* that can become new therapeutic targets.

To address these issues, we have used two lentiviral libraries (Cellecta Inc.): i) a clonal tracking library composed by 30 million different barcodes; ii) an shRNA library composed by ~1000 different shRNA able to silence ~100 genes. The libraries were used to infect xenotransplants of patient-derived leukemias and murine leukemias obtained from mouse models. The infected blasts were transplanted in recipient mice and, following leukemia development *in vivo*, the DNA extracted from infected blasts has been used, through massive sequencing techniques, to investigate leukemia growth and evolution (clonal tracking library) and genes required for AML growth (shRNA screening).

Our clonal tracking analysis highlighted a strong clonal selection troughout leukemia evolution with the disappearance even of clones that were present at very high frequency within the tumor population. Nonetheless, the clonal composition of all tumors appeared to be very similar in terms of number of clones, of their frequency distributions and even of

their identity. Taken all together, our data strongly support a model in which each LIC is endowed with an intrinsic and highly variable replicative potential, with the vast majority of LICs that are not immortal and do not possess an indefinite self-renewal ability As a consequence, each leukemia possesses a definite and small number of LICs able to propagate the disease long-term. These LICs tend to overgrow all other clones and to constitute a very large proportion of the tumor.

Concerning the shRNA screening, we believe we set up an innovative *in vivo* approach that allowed us to identify putative therapeutic targets involved in the growth of primary tumors in their natural environment and that is mainly selective for targeting the CSC population. Most of the identified potential targets have been already reported to play a fundamental role in different types of cancer, including leukemia. Because we were able to perform our screening on several AMLs and identified both tumor-specific and tumor-wide putative targets, our study might offer solutions of therapeutic intervention in a wide group of AML patients, hopefully increasing their survival and decreasing therapy related side effects.

1 INTRODUCTION

1.1 Origin and development of blood cells: hematopoiesis

The blood contains several different cell types. Each of these cell types is quite distinct in appearance, and each has a specific biological function. Despite this extreme structural and functional heterogeneity, strong evidence exists that all the blood cells derive from a single cell type: the hematopoietic stem cell (HSC). The processes involved in production of all the various cells of the blood from the HSCs are called *hematopoiesis*. Hematopoiesis, that begins early during embryogenesis and undergoes many changes during fetal and neonatal development, includes HSC self-renewal, HSC commitment to specific lineages, and maturation of lineage-committed progenitors into functional blood cells. The mature blood cells have finite life spans, which vary depending on the cell type. For this reason, during the lifetime of an individual, mature cells are constantly replaced as the oldest cells are removed and the new one formed. Because mature blood cells are predominantly short lived, hematopoietic stem cells (HCSs) are required to replenish the multilineage progenitors and the precursors committed to individual hematopoietic lineages[1].

The existence of HSCs and their capacity for extensive self-renewal have been demonstrated by animal reconstitution experiments with murine hematopoietic cells that were individually genetically marked. In these pioneering experiments, the authors infected blood cells *in vitro* with a recombinant retrovirus able to integrate randomly into the genomic DNA of the host cell, providing a specific marker to track the progeny of those cells *in vivo*. Each unique integration allows to distinguish and to identify the progeny originated from each infected cell based on the size of the bands derived from the genomic DNA after digestion with a restriction enzyme that does not cut the provirus. In the recipient mice of these marked cells, it was observed that, even several months after transplantation, all types of cells in the blood and lymphoid organs derive from an individually marked cell, proving that the cell of origin was multipotent. Moreover, these clones of marked cells continued to contribute to all the hematopoietic lineages for an extended period of more than 8 weeks and also following serial transplantation in

secondary and tertiary recipients. These experiments demonstrated that these cells are able to reproduce themselves (self-renew) over a long period of time and that long-term reconstitution of both the myeloid and lymphoid compartments can be achieved by transplantation of a single murine HSC, indicating that a single HSC is the smallest repopulating unit in mouse[2] .

However, in the HSC compartment, the stem cells are heterogeneous with respect to their ability to self-renew. The actual hierarchical model of hematopoiesis suggests that the HSC pool consists of at least 2 functionally distinct subpopulations: long-term and short-term repopulating HSCs (LT-HSCs and ST-HSCs, respectively). The LT-HSCs have extensive (life-long) self-renewing potential and on commitment give rise to ST-HSCs with more restricted self-renewing capacity. The LT-HSCs are the only cells able, following transplantation procedures in lethally myeloablated recipients, to give to the totality of blood cells for more than 8 weeks [2] .

Transplantation assays in limiting dilutions, up to single-cell, in combination with the isolation of cell subpopulations based on the expression of unique cell surface markers profiles have allowed the isolation of at least another class of multipotent cells with short term (ST-HSCs) reconstitutive potential. Indeed, it has been shown that ST-HSCs can differentiate into all types of blood cells, however, they have a more limited self-renewal capacity compared to LT-HSCs and are able to give rise to the blood lineages for only 8–12 weeks. Recently, Dick and colleagues demonstrated that transplantation of a single human $CD34^{+}$ $CD38^{-}$ $CD45RA^{-}$ $Thy1^{+}$ Rho^{lo} $CD49f^{+}$ cell into immunocompromised mice provides multilineage reconstitution, indicating that also in humans the HSC is the smallest repopulating unit[3].

The process of hematopoiesis can be described as hierarchical, with at the top the LT-HSCs, that self-renew for all the lifetime of an individual, and originate ST-HSCs. These cells progress to multipotent progenitor cells (MPP), an intermediate progenitor stage that commits to one of the two oligo-lineage restricted cells: the common lymphoid progenitors

(CLPs) and the common myeloid progenitors (CMPs). CLPs are restricted to give rise to T-lymphocytes, B-lymphocytes and Natural Killer (NK) cells whereas CMPs are progenitors for the myeloerythroid lineage. Therefore, the CMPs give rise to megakaryocyte-erythrocytes progenitors (MEPs) and granulocytes-macrophage progenitors (GMPs). Finally, these progenitors provide the entire lineage-committed effector cells of the blood system. In physiological situation, each stage of differentiation of the multipotent cells involves functionally irreversible maturation steps and no dedifferentiation or transdifferentiation is permitted between haemato-lymphoid progenitors[4] (Figure 1).

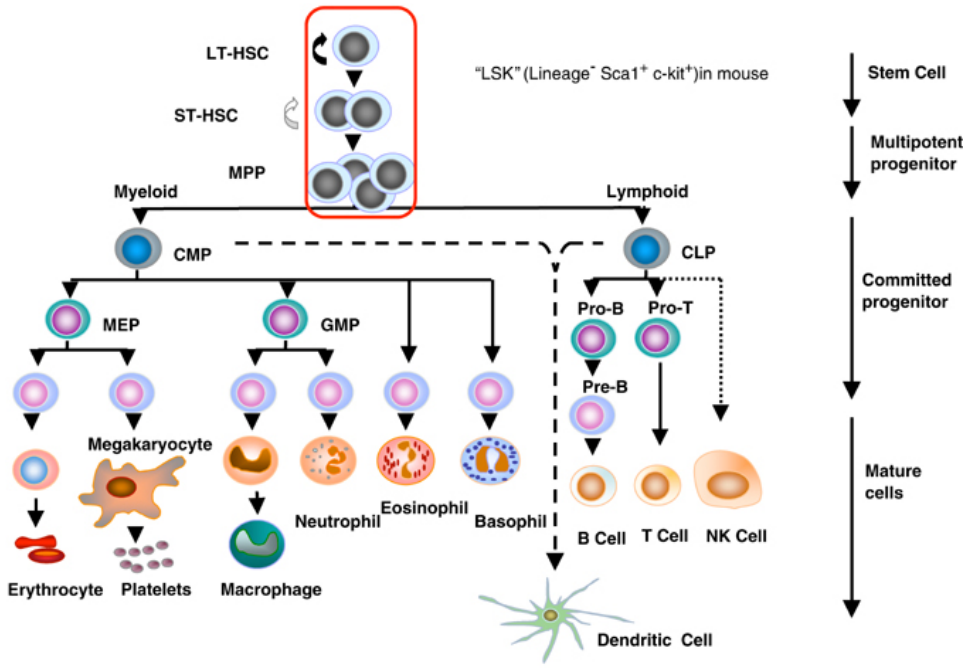

Figure 1. **The hierarchy of the hematopoietic system.** LT-HSC, long-term repopulating HSC; ST-HSC, short-term repopulating HSC; MPP, multipotent progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; MEP, megakaryocyte/erythroid progenitor; GMP, granulocyte–macrophage progenitor. The encircled pluripotent populations, LT-HSC, ST-HSC and MPP, are characterized by a surface markers profile defined as Lin-, Sca-1+, c-kit+ (LSK)[5].

1.2 Hematopoietic stem cells and cancer stem cells

As already described, HSCs are the only cells able to maintain haematopoiesis by their multipotency and self-renewal potential[6]. The HSCs are present in the bone marrow in about 0.01–0.2% of the total mononuclear cells. They are largely quiescent, and this allows them to preserve their genomic integrity. Indeed, frequent cell cycle and consequential DNA replications can introduce DNA mutations inducing oncogenic transformation[7]. This quiescent state also protects the HSCs from exhaustion and their activation occurs under stressful situations, such as blood loss or infections[8,9].

Self-renewal is the key biological property of stem cells. It allows each stem cell, after cell cycle and cytodieresis, to produce one or two daughter stem cells through asymmetric or symmetric division, respectively. In any case the daughter stem cells retain the self-renewal capacity, ensuring that the stem cell population is maintained or expanded for long-term clonal growth. The HSCs reside in specific niches in the bone marrow and their homeostasis is regulated by a complex interplay of factors, such as interleukins and chemokines that govern HSCs fate decisions[6].

As for other tumours, also for AMLs, it has been possible to show by xenotransplantation models the existence of cancer stem cells responsible for driving and sustaining the disease: these cells have been called leukaemia-initiating cells (LICs)[10]. Such as HSCs, LICs also stay at the apex of a leukemic differentiation hierarchy and give to all leukemic cells. LICs and HSCs share many functional and molecular characteristics. Indeed, LICs may be considered as the malicious and functional counterpart of HSCs in the hematopoietic tumours. Due to their self-renewal ability, LICs maintain the disease producing the so called blasts, which are highly proliferative, have a block in terminal differentiation and defect in apoptotic mechanisms[11,12]. Moreover, proprieties like quiescence, increased efflux pump activity and localization in distinct bone marrow niches, similar to the one that characterize the normal HSCs, make the LICs rather resistant to various standard therapies[13,14]. However, these common characteristics should not imply

that the LICs necessarily originate from HSCs that acquire the first transforming mutations. In fact, the use of several murine leukemia models has yielded conflicting results regarding the identity of the cells of origin of leukemias. While some fusion oncogenes, such as MLL-GAS7, seem to be able to initiate leukemias only when expressed in multipotent progenitors (MPP)[15], others, such as PML-RAR, have more mature myeloid committed progenitors as targets for leukemic transformation (Figure 2).

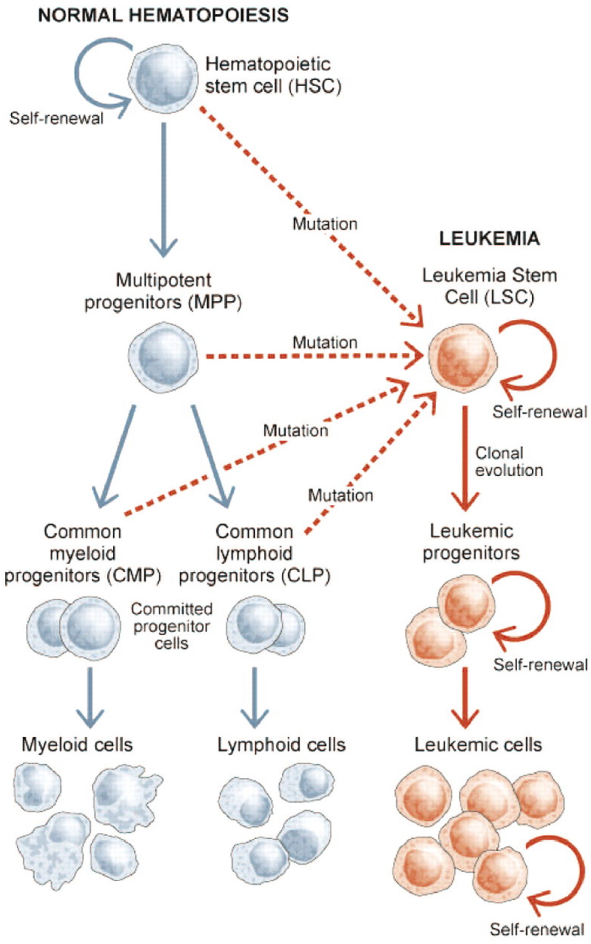

Figure 2. The leukemia-initiating stem cell model. Functional studies on acute myeloid leukemia have led to the identification and characterization of the leukemia initiating cells (LICs), which share similar biological features with the normal hematopoietic stem cells (HSCs). LICs can derive either from HSCs, but also from other type of cells such as multipotent progenitors (MPPs), committed progenitor cells or even more differentiated cells that accumulate mutations that restore the stem cell features required to generate and regenerate tumors[16].

Moreover, heterogeneity in terms of self renewal potential has been observed even within the same tumour. Indeed Hope et al. retrovirally transduced primary human AML cells

with a GFP construct to track the leukemic cells *in vivo*, through its integration site, after transplantation into NOD-Scid mice[10]. This clonal evolution study revealed that the clonal population of cells initially populating host mice was disappearing after approximately 12 weeks and being replaced by a population not detected earlier. This findings show that LICs are not functionally homogeneous but, like the normal HSCs, comprise distinct hierarchical classes. Analysis of the clonal integration sites in secondary and tertiary recipients revealed that often clones detectable in the primary mouse were also present in the secondary mouse, indicating self-renewal of this clonal population in the primary mouse. Interestingly, as for their normal counterpart, some clones were contributing to leukemia engraftment persistently for >12 weeks in the primary recipients but consistently failed to engraft for more than 12 weeks in secondary and tertiary recipients. This can be considered as evidence that short-term LICs arise from long-term LICs, which are believed to be quiescent and likely persistent below the detection limit in the primary recipients[10].

1.3 Leukemia

Leukemia is defined as a cancer that affects the bone marrow and results in abnormal proliferation of white blood cells, also called leukocytes. Different types of leukemia exist and they can be classified according to the type of blood cell that is affected and to how quickly the disease develops and progresses.

According to the time of evolution of the disease, leukemias can be classified in acute and chronic. In acute leukemias, the blasts remain in a immature state with the impossibility to carry out their normal functions. In this type of leukemia, the blasts increase rapidly in number with a consequent rapid worsening of the disease. In chronic leukemia, the blasts are more mature and maintain the ability to carry out some of their normal functions. In this case, the increase in blasts number is less rapid and the development of the disease slower compared to the acute diseases[17].

Depending on which lineage is affected, leukemia can be subdivided in two main categories: myeloid and lymphoid. In the myeloid leukemia cells of the myeloid lineage such as granulocytes, monocytes, erythrocytes or megakaryocytes are involved. In the lymphoid leukemia cells of lymphoid lineage such as B-cells, T-cells or natural killer (NK)-cells are involved[17].

Based on the two parameters described above, it is possible to classify the leukemias in four main categories: acute myeloid leukemia (AML); chronic myeloid leukemia (CML); acute lymphoblastic leukemia (ALL) and chronic lymphocytic leukemia (CLL)[18]. The studies presented in my thesis are performed exclusively on samples of AML, therefore I will focus my attention and describe in more detail only this type of leukemia.

1.4 AML

Acute myeloid leukemia (AML) is the most frequent haematological malignancy in adults and accounts for ~75% of all acute leukemias, with an estimated worldwide annual incidence of 3-4 cases per 100,000 people[19]. The two main systems used to classify AML into subtypes are the French-American-British (FAB) classification and the more recent World Health Organization (WHO) classification. In particular, the FAB classification was elaborated in the 1970s and is based mainly on morphological appearance of the blasts and their cytochemistry the FAB classifies the AML diseases into categories from M0 to M7, according to the type of cell from which the leukemia develops and the degree of differentiation of the leukemic cells[20] (Table 1).

Table 1. Fab classification of Leukemia.

FAB subtype	Description	Comment
M0	Undifferentiated	Myeloperoxidase negative, myeloid markers positive
M1	Myeloblastic without maturation	Some evidence of granulocytic differentiation
M2	Myeloblastic with maturation	Maturation at or beyond the promyelocytic stage of differentiation
M3	Promyelocytic	Also called APL.
M4	Myelomonocytic	Characterized from granulocytic and monocytic blast
M4eo	Myelomonocytic with bone marrow eosinophilia	Characterized by inversion or translocations of chromosome 16
M5	Monocytic	M5a without maturation or M5b with partial maturation
M6	Erythroleukemia	Erythroblast >50% of nucleated cells, myeloblast > 20% of nonerythroid cells
M7	Megakaryoblastic	AMLs associated with Down's syndrome

Even if clinically useful, the FAB classification system has one major limit, it does not take into account many factors that are now known to affect prognosis.

For these reasons, in 2008 the World Health Organization (WHO) classified AML by the presence of recurrent chromosomal rearrangements, which have significant prognostic values (Table 2). Indeed, based on their cytogenetic features, AML patients can be broadly classified into three risk groups according to their response to standard therapy: favourable risk for patients harbouring t(15;17), t(8;21) and inv(16); intermediate risk for patients with normal karyotype (NK) or t(9;11); adverse risk for patients with complex karyotype (harbouring 3 or more abnormalities), del(7), del(5), inv(3)/t(3;3)[21].

Table 2. WHO classification of Leukemia.

Acute myeloid leukemia with recurrent genetic abnormalities
AML with t(8;21)(q22;q22); RUNX1-RUNX1T1
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); CBFB-MYH11a
APL with t(15;17)(q22;q12); PML-RARA
AML with t(9;11)(p22;q23); MLLT3-MLL
AML with t(6;9)(p23;q34); DEK-NUP214
AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1-EVI1
AML (megakaryoblastic) with t(1;22)(p13;q13); RBM15-MKL1
Provisional entity: AML with mutated NPM1
Provisional entity: AML with mutated CEBPA
Acute myeloid leukemia with myelodysplasia-related changes
Therapy-related myeloid neoplasms
Acute myeloid leukemia, not otherwise specified
AML with minimal differentiation
AML without maturation
AML with maturation
Acute myelomonocytic leukemia
Acute monoblastic/monocytic leukemia
Acute erythroid leukemia
Pure erythroid leukemia
Erythroleukemia, erythroid/myeloid
Acute megakaryoblastic leukemia
Acute basophilic leukemia
Acute panmyelosis with myelofibrosis
Myeloid sarcoma
Myeloid proliferations related to Down syndrome
Transient abnormal myelopoiesis
Myeloid leukemia associated with Down syndrome
Blastic plasmacytoid dendritic cell neoplasm

Despite intensive search for new therapies and prognostic markers, AML is still a disease characterized by highly variable prognosis among patients and high mortality rate. Indeed, the overall survival rate (OS) of adult AML patients at 5 years is less than 50% and, in the "oldest people" only 20% survive two years[22]. The gold-standard treatment of AMLs, independently from their clinical and morphological classification, is based on chemotherapy, consisting of an induction therapy with 3 days of anthracycline and 7 days of cytarabine ("3+7", or approaches of similar intensity), followed by diversified consolidation strategies. Standard treatments induce complete remission (CR) in up to 80% of AML patients. However, ~70% of them relapse within 5 years, and eventually succumb

to the disease[23]. The relapse of AML is thought to result from residual chemoresistant leukemic cells still present at CR, although below the limits of detection of conventional morphologic and molecular assessments.

An exception to this situation are the myeloid leukemias harbouring the t(15;17), defined as APLs, that results in expression of the fusion protein PML-RAR, for which the standard treatment is based on a differentiation therapy induced by trans-retinoic acid (ATRA). This is one of the first examples of differentiation therapy developed following the dissection and unravelling of the molecular functions of the product that derives from the fusion of the gene coding for the promyelocytic protein (PML) with the retinoic acid receptor-alpha (RARa). The fusion product blocks the activity of the PML protein by recruitment of co-repressors and histone deacetylases that prevent transcription of genes involved in cellular differentiation and, therefore, block the cells in an immature state. Pharmacological doses of ATRA can bind the RARa component of the fusion product, inducing its degradation by ubiquitinization. This allows for the normal functions of PML, promoting granule formation and differentiation of the promyelocytes[24].

In this example, specific inhibition of the genetically altered activation molecules of the leukemic cells allows the leukemic cells to differentiate and die. Because acute myeloid leukemias usually have mutations of more than one gene, combinations of specific inhibitors that act on the effects of different specific genetic lesions promises to result in more effective and permanent treatment. However, we still need to clarify the leukemogenic process and to identify the genetic lesions and molecular mechanisms that contribute to the development of AML in order to be able to carry out a more accurate risk stratification of newly diagnosed patients and to identify critical targets for the design of patient-specific treatments.

1.5 Multiple hit model of AML

AML is a multistep disease. The primary lesions are necessary but not sufficient for the full development of disease, which is observed only following the acquisition of cooperative mutational events[25]. This principle has been demonstrated in mouse models, where introduction of the fusion gene found in acute leukemia by transgenic technology or retroviral transduction results in predisposition to acute leukemia with long latency, unless the accumulation of additional genetic hits is facilitated by treatment with a mutagenic agent. This multistep carcinogenesis induces cancer cells to acquire all the properties in order to become malignant: sustaining proliferative signaling, evading growth suppression, enabling replicative immortality, resisting cell death, inducing angiogenesis, activating invasion and metastasis, avoiding immune destruction, and deregulating cellular energetics[26].

It has been proposed that this multistep process requires the synergy and the cooperation of two classes of mutations: class I mutations that activate signal transduction pathways and confer a proliferation advantage on hematopoietic cells, and class II mutations that affect transcription factors and primarily serve to impair hematopoietic differentiation[27-29] (Figure 3).

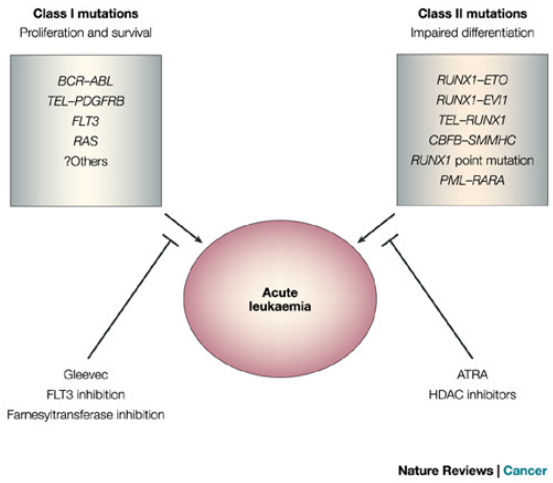

Figure 3. Multiple hit model in AML. According to this theory two classes of mutations are needed. Class I mutations affect genes that are involved in signal transduction and lead to increased cell proliferation and/or survival and class II mutations that impair differentiation[30].

The class I mutations can increase proliferation and survival potential of both hematopoietic stem and progenitor cells, often leading the constitutive activation of protein tyrosine kinases (PTKs) normally without involving cellular differentiation. Class I mutation can lead both to gain-of-function or loss-of-function of the protein. Examples of proteins constitutively activated by gain-of-function mutations are: ABL, JAK2, FLT3, PDGFR, KIT and activating mutation of the RAS family members[25,31-34], while, examples of proteins that lose their activity by loss-of-function alterations are NF1 or PTPN11[35,36].

The class II mutations affect differentiation and are able to increase self-renewal properties of hematopoietic cells. This kind of mutation usually can works on chromatin architecture, recruiting either co-activator or repressor complexes, while having modest roles on cell proliferation or survival. Examples of this class of mutations comprise gene rearrangements that act modifying transcriptional regulation on target such as the core-binding factor CBF, CEBP, RAR, and MLL[37,38], and components of the transcriptional activation complex such as CBP, MOZ, and TIF2[39], which play a crucial role during the differentiation of hematopoietic stem and progenitor cells. Are considered Class II mutations all the chromosomal aberrations such as t(8; 21), inv(16) and t(15; 17), which generate fusion transcripts of RUNX1/ETO, CBFβ/MYH11 and PML/RARα[27], respectively.

Not only chromosomal abnormalities but also mutations of the transcription factors RUNX1, C/EBPα and MLL are classified into this group[27].

The multiple hit model is supported by several clinical and experimental data and mutations from different complementation groups were frequently found in the same patient. Recent studies and technologies, in fact, demonstrates that in leukemic cells more than one recurring mutations is present, suggesting, once again, a multi-hit model for leukemogenesis[40]. Studies on fusion genes *MLL/AF4* or *TEL/AML1* have shown that, although some class II mutations arise *in utero*, the development of frank leukemia occurs

with a protracted postnatal latency, indicating the need of additional secondary genetic alteration[41].

In the acute promyelocytic leukemia (APL), characterized by the t(15;17) rearrangement, it has been observed that 30% of the patients also harbor a FLT3 activating mutation, suggesting a possible cooperation between these two mutations in inducing APL[42]. This hypothesis has been confirmed in an experimental animal model of APL that provided strong evidence for functional cooperation between the PML/RAR fusion and the activating FLT3 mutation. Briefly, the expression of PML/RAR from the cathepsin G promoter in transgenic mice resulted in an APL-like disease with 6 months latency and incomplete penetrance of ~15%–30%[43]. Co-expression of the FLT3-ITD mutation in the bone marrow of PML/RAR transgenic mice induced APL with a shorter latency and complete penetrance[42]. In particular, FLT3 mutations are considered the "ideal" second hit cooperative mutations as shown in different animal models of AML[44,45].

1.6 Clonal evolution of Leukemia

Clonal evolution and subclonal diversity are hallmarks of the pathophysiologic mechanisms of many cancers including Acute myeloid Leukemia. A successive acquisition of the multiple cancer hallmarks depends on a succession of alterations in the genomes of neoplastic cells and dominates the clonal evolution. This clonal selection is an evolutionary process that is driven by selection and expansion of adapted subclones that acquire proliferative advantage compared with the others[46]. The clonal evolution model of cancer development involves gain of function of oncogenes and loss of function of tumor suppressor genes that cooperate to induce fulminant disease.

This intratumoral genetic heterogeneity has been initially provided by various techniques: chromosome karyotyping[47], genetic analysis of multifocal cancers[27], fluorescence in situ hybridization (FISH)-based tissue section screening[48], immunophenotype-based cell

analysis[49], molecular probing of multiple small biopsies, microselected tissues, and sector ploidy profiling[50].

Recent techniques, such as next generation deep parallel sequencing (NGS) and copy number alteration (CNA) analysis have provided an important improvement in the understanding of subclonal tumor composition and evolution leading to the definition of driver and passenger mutations. Not all the mutation, in fact, contribute equally to tumor evolution. The mutation that confer a selective growth advantage and promote clonal expansion are definited driver mutation. The passenger mutation, instead, are inert mutation and are found within cancer genomes because somatic mutations without functional consequences can occur during cell division. These inert mutations, will propagate through the clonal expansion of tumor and will be present in all cells of the final tumor.

All these recent data have provided strong evidences supporting the concept that tumors diplsay an intrinsic genetic landscape with an highly subclonal heterogeneity and intraclonal genetic diversity [51,52]. Moreover, in some types of tumor, such us breast or pancreatic cancer, the degree of genetic diversification has been linked with poorer prognosis[53,54].

In AML, it has been shown that each patient contains different leukemic subclones that evolve along a complex, branching multi-clonal evolution. In these cases, a shift in clonal dominance can occur and minor clones might become able to outcompete dominant clones[55-57]. Clonal evolution of leukaemia could be described in two principal ways: linear or branched. The first case is represented by an essentially linear architecture with stepwise accumulation of driver mutations (Figure 4). Observing the mutational complexity, by deep sequencing methods, it was possible to show that during the development of cancers, including AML, dominance of subclones can change throughout the course of disease[54,55,57-60], and the clones that contain the highest numbers of lesions are not necessarily numerically dominant within the tumour[40,55,56].

Even if cancers can be traced back to a single cell, the continuous acquisition of lesions dramatically increases genetic and clonal heterogeneity. For these reasons it is likely that most cancers evolve with a complex branching architecture (Figure 4).

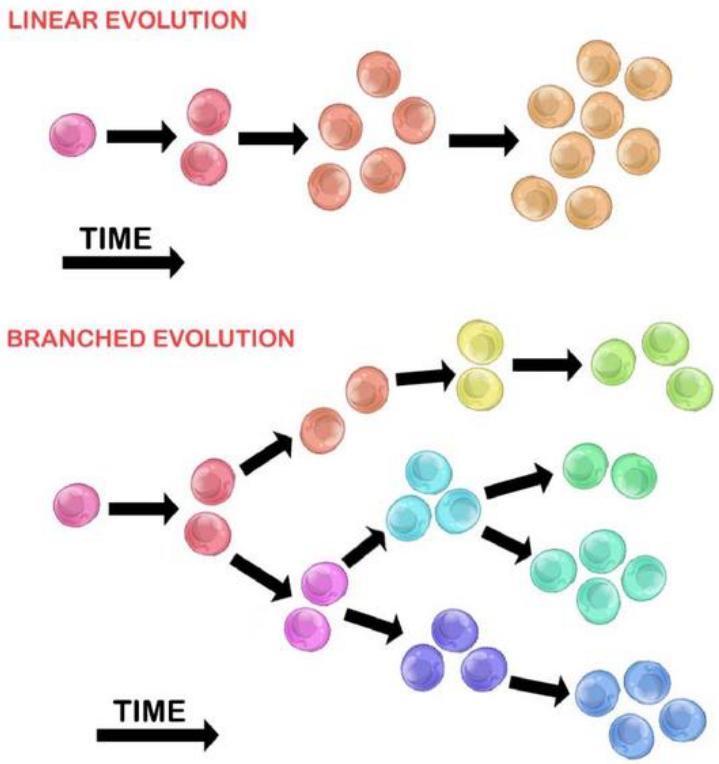

Figure 4. Linear and branching clonal evolution. (A) Linear evolution. The dominant clones are the results of stepwise accumulation of driver mutation. In this manner the final tumor contains all the mutaions acquired during the evolution of desease. (B) Branching evolution. The final composition of leukemia is composed from several clones, each of which contain its subset of specific driver mutation.

Even if the evolution of cancer occurs through a complex branching pattern of mutation acquisition[61], there are evidences that both linear and branched architecture can be present in individual AMLs. A study on 8 primary and relapsed AML samples shows that in some cases (3/8), only a single mutation cluster was found in the primary tumour. In these cases, the single clone gained additional mutations at relapse, consistent with a linear pattern of evolution, although minor branching subclones could have been missed (Figure 4). In the remaining case (5/8), multiple mutation clusters corresponding to different subclones were detected in the primary sample[57] (Figure 5A).

Other two studies confirmed that re-emergence or evolution of a founder or ancestral clone is typical in relapsed AML[62,63].

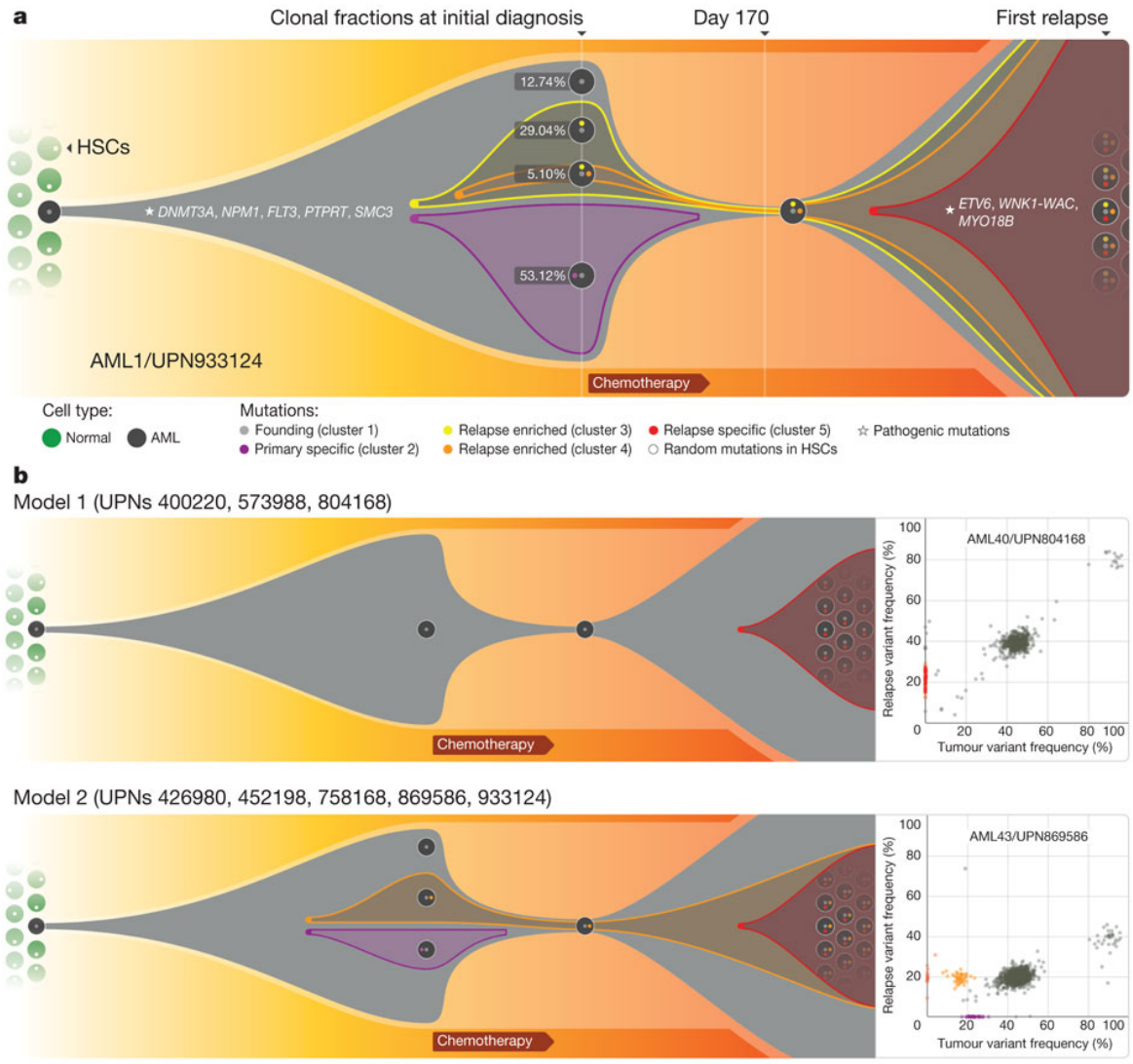

Figure 5. Representation of clonal evolution from the primary tumour to relapse. A) The founding clone in the primary leukemia contain all the somatic mutations that are recurrent in all AML cells: one subclone within the founding clone evolved to become the dominant clone at relapse by acquiring additional mutations. **b**, Examples of the two major patterns of tumour evolution in AML. Model 1 shows the dominant clone in the primary tumour evolving into the relapse clone by gaining relapse-specific mutations; Model 2 shows a minor clone carrying the vast majority of the primary tumour mutations survived and expanded at relapse[57].

1.7 Chemo-resistance

The mechanisms of chemo-resistance of AMLs are currently unknown, however, by whole-genome sequencing of primary/relapse tumor pairs, the relapse-specific mutations have been demostrated. These mutations were detectable in the primary tumors only in

some of the patients analysed. A central open question in cancer treatment is what is the precise nature of the interaction of clonal evolution with cancer therapy.

Initial studies explained the presence of relapse-specific mutation with the potential mutagenic role of chemotherapy (Figure 6A). Studies of WGS (whole genome sequencing) on Acute myeloid malignancies have suggested that the novel mutagenesis may result from the genotoxic effects of chemotherapy. This observation was supported also by a changing spectrum of somatic single-nucleotide alterations[64]. A different situation has been shown in the case of Chronic Lymphoid Leukemia (CLL), where the contribution of the chemotherapy's mutagenizing effect seems to be limited[65]. Therefore, although chemotherapy-induced mutagenesis has the potential to contribute to further clonal diversification, other sources for generating evolutionary shifts appear to be at play, and probably involve pre-existing genetic variants or subclones[57]. As shown in Figure 6 two models have been proposed to explain chemoresistance and AML relapse: i) in the first model proposed, the chemoterapy can induce novel mutagenesis that confer resistance to a clone that became dominant (Figure 6A). ii) a minor resistant leukemic clone survive to chemotherapy and expand at relapse (Figure 6B). In this model the selection and expansion of resistant clones is induced by the selective pressure of therapy that is often directed at a particular genetic context which may not be shared by all subclones. In this case, no new mutation occurs, but a minor clone already contain all the features that confer resistance[57,64].

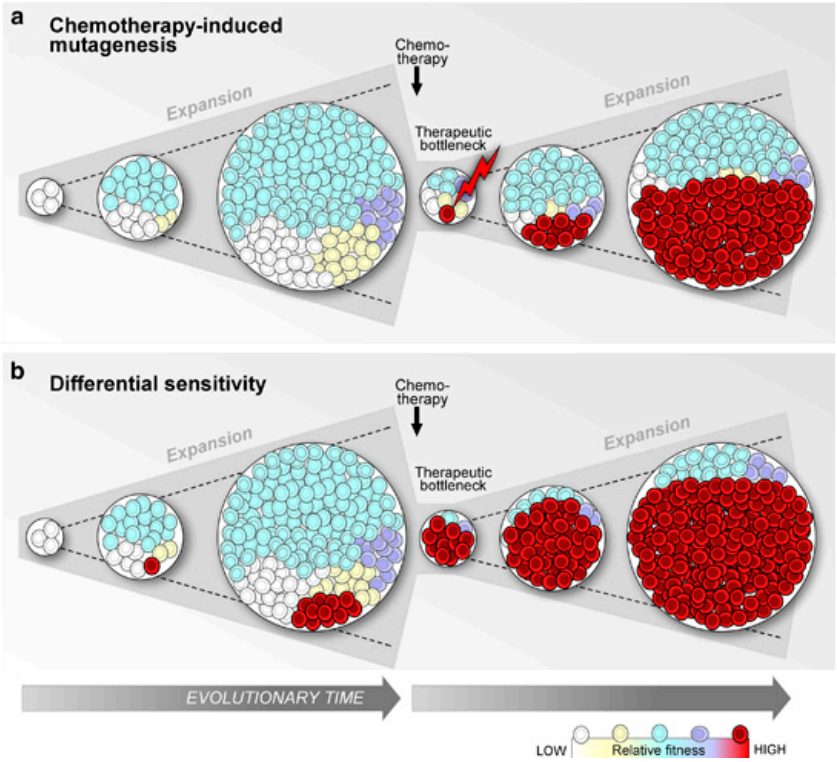

Figure 6. Two models of how cancer therapy may accelerate clonal evolution. a) in the first model, cancer therapy containing genotoxic agents can induce novel mutagenesis leading to resistance. b) in the second model, no new mutations are acquired but therapy can accelerate clonal evolution by selecting a clone that already contain a mutation that confers the resistance to the therapeutic agent[64].

1.8 Leukemia Xenograft model

Even the most clinically aggressive sample of leukemia cells from a patient is unlikely to grow when placed in culture, and cell characteristics can change depending on the media used[66-68]. Additionally, culture systems cannot replicate the interactions between the leukemia cells and their stroma and immune cells. In other words, it is impossible to reproduce *in vitro* the niche where the leukemic cells growth and expand. To circumvent these limitations, murine xenograft models have been developed to allow engraftment of primary patient samples[66]. Xenografts in immunodeficient mice allow performing functional studies on patient-derived AML cells[69]. The mouse is frequently used as a model due to its relative physiologic similarity to humans and its capacity for facile genetic manipulation, as well as because of the relative efficiency of breeding and housing mice[70]. Although xenograft models have been extensively used in leukemia studies, there are some disadvantages and limitations to the utility of these models. For instance, the abbreviated

life span, increased metabolic rate, and altered telomere length of inbred mouse strains, as well as the altered prevalence and time of onset of cancers are examples of some fundamental physiological differences between humans and mice[71]. In particular, if the immunodeficiency is a prerequisite for preventing human tumor cell rejection, in xenograft models the growth of tumor occurs in a host with an impaired immune system.

Engraftment of patient-derived cells in immuno-compromised mice has been established in the early 1990s using severe combined immunodeficiency (scid) mice. Engraftment capacity has improved by using more severely immuno-compromised mice[72-74]. We currently use the NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ (NSG) mouse model. These mice combine the features of the NOD background (that spontaneously develops type 1 diabetes), the severe combined immune deficiency mutation (scid) and IL2 receptor gamma chain deficiency resulting in multiple defects in both the innate and the adaptive immune system (lacking B, T, and functional NK cells)[75-79]. Moreover, impaired NK function reduced the need for conditioning with radiation prior to leukemia administration[80-82]. This was of importance, since these mice have a tendency to develop spontaneous thymic lymphoma, and have generally a short life span[83]. In NSG mice, engraftment can occur very fast, it is possible to detect blasts in the peripheral blood, and total number of cells required to engraft can be little[75]. NSG mice can even be engrafted with as little as 1–100 cells derived from leukemic patients[84]. The relative simplicity of engrafting human leukemia in these mice made the NSG model excellent for leukemia research.

1.9 shRNA screening

RNA interference (RNAi) facilitates the assessment of gene function by silencing gene expression. It exploits a physiological mechanism that represses gene expression, primarily by causing the degradation of mRNA transcripts. In mammalian cells, physiological RNAi is primarily mediated by non-protein-coding RNA transcripts, known as microRNAs

(miRNAs). miRNAs are produced in a similar manner to mRNAs, but are processed into shorter RNA species containing a hairpin structure, known as short-hairpin RNAs (shRNAs). shRNAs are, in turn, processed into short double-stranded pieces of RNA known as short interfering RNAs (siRNAs). Within the multi-protein RNA-induced silencing complex (RISC), one strand of a siRNA duplex binds a protein-coding mRNA transcript that bears a complementary nucleotide sequence. This interaction allows a nuclease in the RISC to cleave and destroy the protein-coding mRNA, therefore silencing the expression of the gene in a relatively sequence-specific manner [85]. Recently, the use of libraries of shRNAs, based on miRNAs design and cloned in lentiviral vectors, has been shown to be an effective way to perform both *in vitro* and *in vivo* screenings. In particular, this approach *in vivo* has led to the identification of genes involved in the cancerogenesis of some cancer types. An shRNA screening performed in a lymphoma murine model has defined a central role for regulators of actin dynamics and cell motility in lymphoma cell homeostasis impairing lymphoma cell migration to the lymph nodes and other organs that represent common sites of lymphoid metastasis. Validation experiments confirmed that these proteins represent a *bona fide* lymphoma drug targets[86]. In a mouse model of hepatocellular carcinomas a library of pools of shRNAs targeting the mouse orthologs of genes re-currently deleted in a series of human hepatocellular carcinomas has proved their contribution to tumor progression.. Among them, it has been shown that XPO4 encodes a nuclear export protein whose substrate, EIF5A2, is amplified in human tumors and is required for proliferation of XPO4-deficient tumor cells and for the promotion of hepatocellular carcinoma in mice[87]. More recently, both Mllt6 and b-catenin have been identified as top oncogenic screen hits in a model of $Hras^{G12V}$ epidermal hyperplasia and have been shown to play an important role in tumour initiation and maintenance of human squamous cell carcinomas in xenograft models[88]. Concerning AML, up to date, similar screenings have been conducted only on an AML mouse model driven by MLL-AF9 and $Nras^{G12D}$. Several shRNAs targeting Brd4, a member of the BET family of bromodomain-

containing proteins, were among the most strongly depleted, identifying this gene as the top scorer in the screen. These proteins bind to acetylated histones to influence transcription. Knock down or pharmacological inhibition by JQ1 of BRD4 had a profound effect on AML progression *in vivo* in mice transplanted with MLL-AF9/NrasG12D leukaemia cells. A similar effect was observed on the growth of human AML cell lines harboring a variety of mutations besides MLL-AF9[89]. Furthermore, gene expression changes upon BRD4 inhibition were similar to those described by the same researchers upon knock down of Myb in the same AML mouse model. Ablation of Myb expression inhibited the proliferation of leukemic cells *in vitro* and eradicated disease *in vivo* with limited effects on normal hematopoietic cells. Thus, in murine model of AML, Myb appears to be an early and key player in the oncogene addiction mediated by MLL- fusion proteins[90,91] . These data suggest that there are common and overlapping pathways mediating self-renewal downstream of a variety of different initiating mutations that can be identified performing estensive shRNA screenings on AML characterized by different molecular alterations.

2 AIM OF THE PROJECT

AMLs are haematological disorders characterized by the accumulation of a large number of cells blocked at immature stages of differentiation. It has been shown that AMLs are extremely heterogeneous in terms of frequency and origin of cancer stem cells (also called leukemia-initiating cells, LICs). Moreover, AMLs are multi-clonal diseases in which each subclone seems to evolve along a complex and branching multiclonal process. These observations may have important implications for AML treatment. In spite intensive research for new therapies and prognostic markers and the discovery of many cancer genes which are potentially druggable, AML is still a disease with a highly variable prognosis among patients and a high mortality rate[92]. Indeed, the overall survival rate (OS) of adult patients at 5 years is less than 50% and, in the oldest people only 20% survive two years[22]. The relapse is very likely due to the inefficacy of the therapies to target the LICs that, after the first remission, are able to reinitiate the disease. Therefore, it is becoming increasingly evident that many open questions remain about the leukemogenic process and, in most cases, a reliable treatment is still lacking. This is very likely due to the extreme genetic and biological complexity of AMLs, that appear as intrinsic limits to the generation of valid anti-leukemia targeting drugs. Thus, we need new studies that will allow us to understand how blasts grow and evolve during leukemia development and in response to therapies, and to identify appropriate druggable targets for the eradication of the LICs.

Along this line, our study has three main goals: 1) the identification of LICs number through the transplantation of traceable leukemic cells; 2) the study of the clonal evolution of the disease *in vivo* before and after chemotherapy treatment; 3) the identification of gene products involved in maintenance and growth of leukemic blasts and assessment of their usage as novel therapeutic targets. For all aims, our experimental approach is based on the infection of leukemic blasts, both of murine and human origin, with lentiviral libraries and their transplantation in recipient animals.

In order to identify and track each LIC *in vivo*, we planned to couple high-throughput sequencing to viral genetic barcoding, transducing the leukemic blasts with a 30 milion-

barcodes lentiviral library. The incorporation of the barcodes into the genome of the infected cells allows to univocally identify the progeny of each LIC and to follow their destiny after *in vivo* growth. This technique has many important advantages. It allows transplantation of a mixed population and tracking of the cells in their physiological environment. Indeed, it minimizes *in vitro* manipulation, sorting, and transplantation in limiting dilution that can stress the cells causing depletion of the number of LICs and decrease in their leukemogenic potential. It has an increased sensitivity because it allows single cell identification by detection of individual barcodes. Moreover, because it does not rely on restriction digestions, it avoids the problem of differential amplification of fragments with varying lengths or melting temperatures. Therefore, this is an unbiased and quantitative method that provides accurate information about the proportionality of the clones deriving from different LICs. Indeed, based on the number and the relative distribution of barcodes, it allows establishing the number of LICs and their relative contribution to tumor development[93].

In order to identify genes that could be used as therapeutical and druggable targets required for the onset and maintenance of AML, we performed a large-scale *in vivo* shRNA screening on the human AML samples available. By comparing the composition of the shRNAs present in the preinjection pool of cells to the one retrieved following tumour development *in vivo* and by measuring their relative abundance (calculated as the ratio between the number of reads at tumour development divided by the number of reads in the preinjection pool) we will able to define which shRNAs have been enriched and which have been depleted during tumour growth. After tumor development *in vivo*, the positive-selected shRNAs should allow the identification of tumour suppressor genes and the counterselected shRNAs should allow the identification of oncogenes required for AML development.

3 MATERIALS AND METHODS

3.1 Buffers

- PBS 1X: 137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, pH=7.4;
- Red Blood Cell (RBC) Lysis Buffer: 8.125 mg/mL NH_4Cl, 1 mg/mL $KHCO_3$, 0.13 mM EDTA;
- Luria Broth (LB): 2.5g Luria Broth/L
- Terrific Broth (TB): 4.7g Terrific Broth/L
- Annealing buffer 2X: 200 mM CH_3COOK, 60 mM HEPES-KOH, 4 mM Mg-acetate
- 2X HBS: 50 mM HEPES pH7.1, 280 mM NaCl, 1.5 mM Na_2HPO_4
- TAE (Tris-Acetate-EDTA, 50x): 242 g/l Tris base, 57.1 ml/l glacial acetic acid (37.5%), 18.6 g/l EDTA.

3.2 Mammalian cell culture

All cells were maintained at 37°C, 5% CO_2 and 20% O_2.

Phoenix 293T cells were cultured in DMEM, supplemented with 10% Foetal Bovine Serum (FBS, South American), 100 U/mL Penicillin, 100 U/mL Streptomicin, 2 mM L-Glutamine;

NB4 cells were cultured in RPMI, 10% Foetal Bovine Serum (FBS, South American), 100 U/mL Penicillin, 100 U/mL Streptomicin, 2mM L-Glutamine;

AML samples were cultured in RPMI, 20% Foetal Bovine Serum (FBS, North American), 2 mM L-Glutamine, 1% 5637-conditioned supernatant.

3.3 Xenografts of human AMLs

Since human acute myeloid leukemia (hAML) xenotransplanted in immunocompromised animals retain the phenotypic characteristics of the original sample, we established some xenografts of hAMLs. We took advantage of the humanized mouse models generated by engraftment of immunodeficient mice with components of the human immune system, the

NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) immunocompromised mice, as the host. NSG mice have been shown to support greater engraftment of human hematopoietic stem cells (hCD34+ cells) than all other strains[75].

Bone marrow or peripheral blood mononuclear cells freshly collected from AML patients are CD3-depleted using the CD3 MicroBeads kit from Milteny Biotech. The patients enrolled in the study gave written informed consent. The human samples are characterized for their karyotype and, by Sanger sequencing, for two of the most recurrent mutations in AMLs: mutations in FLT3 and in NPM1.

Our xenotransplantation protocol consists in the intravenous injection from 5 x 10^5 to 12.5 x 10^6 cells of CD3-depleted mononuclear cells freshly collected from the AML patients in 6-8 weeks old NSG mice. Our rate of success is around 70%.

The level of engraftment of the human blasts in the transplanted animals is determined by FACS analysis of the peripheral blood (PB) cells stained with antibody specific for the human CD45 hematopoietic surface antigen (hCD45) and according to the protocol described in the section 3.11.

3.4 Animal Studies

The PML-RAR knock-in mice were provided by T.J. Ley[43] and backcrossed in the C57BL/6J background. This model expresses the fusion protein PML-RAR in early myeloid cells under the control of human cathepsin G regulatory sequences. Acute promyelocytic leukemias develop spontaneously in this KI model in ~70% of animals between 6 and 16 months of age (average latency of ~1 year) [43].

All mice used in the study for transplantation of the leukemic blasts were purchased from Charles River: C57BL6/Ly5.2 (strain C57BL/6NCrl), C57BL6/Ly5.1 (strain B6.SJL-PtprcaPepcb/BoyCrl), NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ (NSG). The NSG is the most immunodeficient mouse strain[94]. NSG mice lack mature T cells, B cells, and natural

killer (NK) cells[83]. This model preserves microenvironment during long-term engraftment[95] allowing the study of primary human tumor.

3.5 Vectors

- **pRSI16cb-U6-13kCB18-13kCB18-HTS6-UbiC-TagRFP-2A-Puro:** Lentiviral vector that contains two barcodes of 18bp cloned under the promoter U6. The vector contains genes that encode for puromycin and ampicillin resistance and the reporter gene TagRFP. Puromycin resistance and TagRFP are constitutively expressed under the control of UbiC promoter. This vector was used in the clonal tracking experiments (Figure 7).

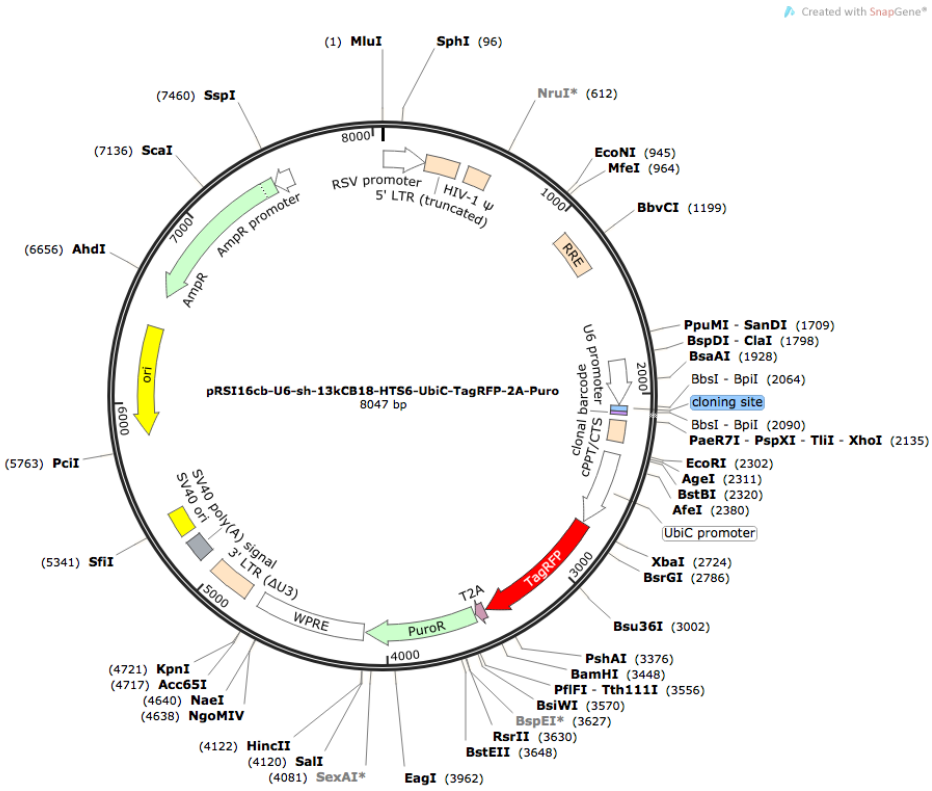

Figure 7. Map of the lentiviral vector used for the clonal tracking experiments: pRSI16cb-U6-13kCB18-13kCB18-HTS6-UbiC-TagRFP-2A-Puro.

- **pRSI17cb-U6-sh-13kCB18-HTS6-UbiC-TagGFP2-2A-Puro:** Lentiviral vector that allows constitutive expression of the shRNAs under the U6 promoter. To each shRNA is associated a unique barcode of 22 bp and a variable number (up to 13,000 possible

combinations) of clonal tracking barcodes of 18 bp. Differently from the clonal tracking vector, it expresses the TagGFP reporter gene. This vector was used in the shRNA screening experiments (Figure 8).

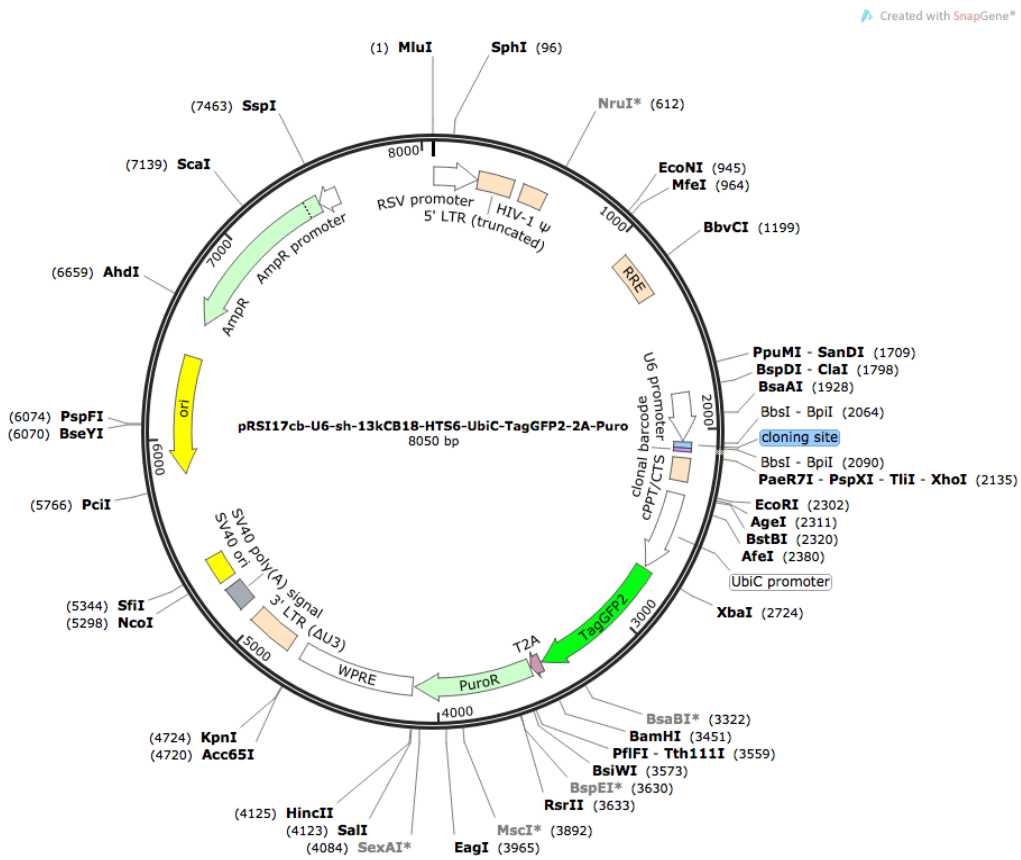

Figure 8. Map of the lentiviral vector used for the shRNA screening: pRSI17cb-U6-sh-13kCB18-HTS6-UbiC-TagGFP2-2A-Puro.

- **pINDUCER11 (miR-RUG):** inducible lentiviral vector that expresses constitutively rtTA3 and eGFP under the hUBc promoter. Doxycycline (dox) administration allows the expression of TagRFP and shRNAs, cloned under a tetracycline dependent promoter TRES2 (tetracycline response element 2). This vector was used for validation of the target shRNAs identified by our screening (Figure 9).

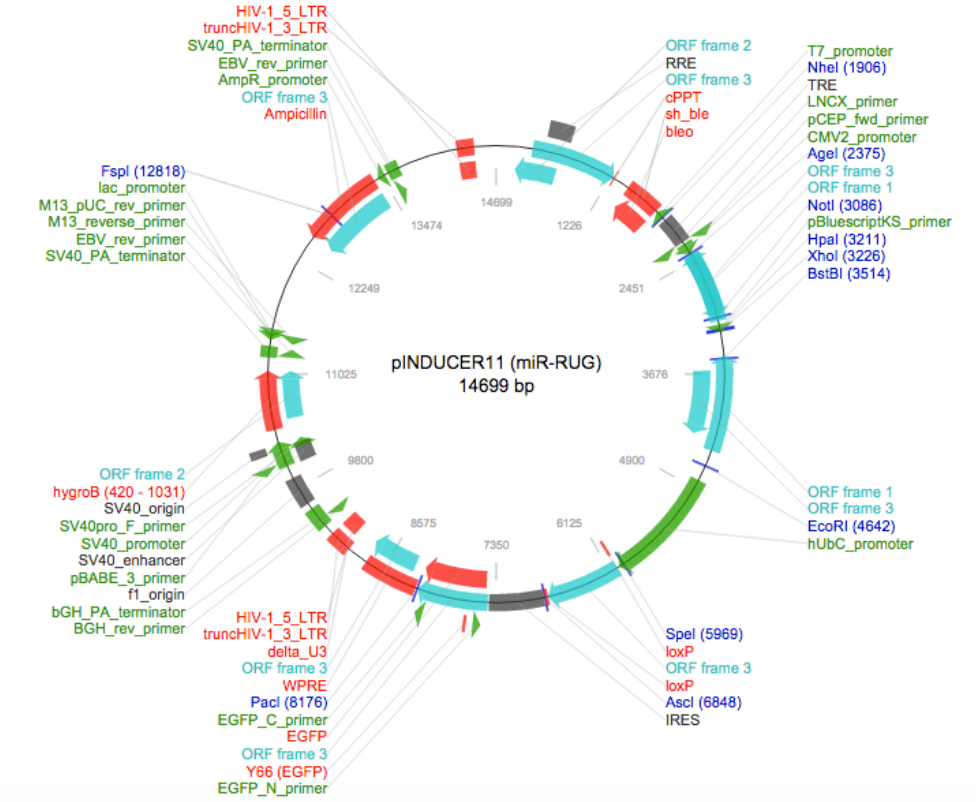

Figure 9. Map of the pINDUCER11 (miR-RUG) inducible vector.

3.6 Lentiviral packaging vector

- **pMDLg/pRRE:** packaging lentiviral vector derived from HIV-1. It express for HIV-1 genes gag and pol. In general, all the vector HIV-1-derived are deficient for all the genes (vif, vpu, vpr nef) that contribute to viral pathogenicity. Is used in lentiviral production with packaging vector of 3rd generation in combination with pRSV-Rev and pMD2.VSVG.
- **pRSV-Rev:** packaging lentiviral vector that express for HIV-1 Rev shuttle protein. Rev, in fact, allows to viral pre-mRNA to be carry out from the nucleus. It is used in lentiviral production with packaging vector of 3rd generation in combination with pMDLg/pRRE e pMD2.VSVG.
- **pMD2.VSVG:** envelope vector derived from Vescicular Stomatitis Virus. It is used to pseudotype viral particles with VSV-G proteins in lentiviral production with packaging vector of 2nd and 3rd generations.

- **pCMVR8.2:** packaging lentiviral vector HIV-1-derived. It expresses the genes gag, pol, rev, and tat for HIV. It is used in lentiviral production with packaging vector of 2nd generation in combination with pMD2.VSVG.

3.7 Lentiviral libraries

3.7.1 Clonal tracking Library

For the clonal tracking experiments, we purchased from Cellecta a lentiviral library containing exclusively two consecutive 18-nucleotide barcodes cloned after the U6 promoter. Because there is a combination of 13000x13000 different 18-nucleotides barcodes, the complexity of the clonal tracking libraries is of around $30x10^6$ unique barcodes. The lentiviral expression vector described in section 3.5 contains also a puromycin-resistance gene ($Puro^R$) and a red fluorescent protein marker (TagRFP) under control of UbiC promoter.

3.7.2 shRNA screening Library

The shRNA library used in our scrrening was purchased from Cellecta. The library is custom made and contains 1204 shRNAs targeting 118 genes (10 shRNAs *per* gene) involved in epigenetic regulations. The lentiviral expression vector, described in detail in setion 3.5, contains also a puromycin-resistance gene ($Puro^R$) and a green fluorescent protein marker (TagRFP). The shRNAs are expressed under the U6 constitutive promoter and are linked to a unique 18-bp barcode that univocally identifies each shRNA and can be amplified by PCR. Moreover, each shRNA is cloned in a library of vector backbones that contains a second barcode of 18 nucleotides. There are up to 12415 different combinations *per* shRNA of these second barcodes. This characteristic allows for clonal identification of the cells targeted by each shRNA.

3.8 Production of the lentiviral particles

Packaging of the lentiviral vectors was performed using 2nd or 3rd generation packaging vectors with two alternative methods of transient transfection: calcium phosphate ($Ca_3(PO_4)_2$) or Lipofectamine®. All transfections were performed using 293T cells plated in 15-cm plates.

3.8.1 Transfection with calcium phosphate

22µg of lentiviral plasmids were mixed with 9µg of envelope plasmid pMD2.VSVG and with 2nd (13µg of pCMVR8.2) or 3rd (5,5µg pRSV-Rev and 11µg of pMDLg/pRRE) generation packaging vectors. $CaCl_2$ is added to the vectors mix at a final concentration of 0.5 M. This mixture is, then, added drop-wise to bubbling 2X HBS solution. After an incubation of 20 minutes, the DNA precipitates are slowly added to the 293T cells monolayer (cells should be 70-80% confluent). Lentiviral supernatants are harvested 24 and 48 hours after transfection. After centrifugation at 2500 rpm for 5 minutes at RT and filtering through 0,2 µm filter (Nalgene Rapid-Flow), the lentiviral particles are concentrated 1000-fold by ultracentrifuging at 75,000 x g for 2 hours at 16°C with Optima™ L-90K Ultracentrifuge (Bechman Coulter) and stored at -80°C[96].

3.8.2 Transfection with Lipofectamine® (Invitrogen)

We prepare two solutions. In the first solution, 6 µg of lentiviral plasmid of interest, 7 µg of pMD2.VSVG vector, 24 µg of pCMVR8.2 vector, 60 µl of PlusTM Reagent are mixed in 1.2 ml of DMEM without FBS and antibiotics. In the second solution, 90 µl of Lipofectamine™ are added to 1.2 ml of DMEM without FBS and antibiotics. Both solutions are first incubated separately for 15 min and then mixed together. After a further incubation of 15 minutes, the DNA-lipid complexes are slowly added to the 293T cells monolayer (cells should be 70-80% confluent). Collection and concentration of the viral supernatants follow the same protocol described in transfection with calcium phosphate.

3.9 Titration of the lentiviral vectors

The titer of the lentiviral preparations can be determined from a serial dilution of the lentiviral stocks by flow cytometry (FACS) or by quantitative PCR (qPCR). The qPCR allows also to measure the number of integrated lentiviral particles in the genome of the transduced cells. We followed the protocols described in Barde et al, 2010[96]

3.9.1 Titration by FACS

The titration of the production of lentiviral particles is carried out, using 293T cells. 250,000 cells/well are plated in a 12 -well plate in 0.5 mL of medium. We carry out a serial dilution (dilution factor 1:10) of the initial viral stock, starting from a dilution of 10^{-1} up to a dilution of 10^{-5}. With 10 µl for each dilution we infect two wells of cells. 72 hours post-infection the cells are collected, fixed in a 1% formaldehyde solution and analysed by FACS for the percentage of TagRFP- or GFP-positive cells (Becton Dickinson LSRII). The multiplicity of infection (MOI) and titer, reported as transducing units *per* ml (TU/ml), is determined using the following equation:

$$titer\ (\frac{U}{\mu L}) = \frac{\text{Number of target cells (count at day 1)} * [\frac{\%\ positive\ cells}{100}]}{volume\ of\ vector\ (ml)}$$

3.9.2 Titration by qPCR

In order to perform the titration by qPCR, before extraction of the DNA, the cells infected with dilution of the lentiviral stock as described above, are usually kept in culture for at least 96 hours in order to reduce the contamination from the not integrated plasmid. The viral copy number *per* cell is determined by qPCR amplification of a genomic sequence, present in two copies *per* genome (i.e. Albumin, Alb), in parallel with viral sequences present in the lentiviral vectors and integrated in the target cells (GAG or WPRE). The Alb genomic sequence is also cloned in the same lentiviral backbone to prepare a standard

curve, used as a reference to count the number of molecules present in the samples. Each sample is amplified using both the Alb and the GAG or the WPRE primers. After amplification the samples are analysed using SDS software (Applied Biosystems) that allows calculation of the standard curve and determination of the GAG or WPRE and Alb quantity for each sample. The copy number *per* cell for each sample is calculated according to the following formula:

$$Vector\ copy\ number = \frac{Quantity\ mean\ of\ WPRE\ or\ Gag\ sequence}{Quantity\ mean\ of\ Alb\ sequence} * 2$$

The sequences of the primers used for the qPCR are listed in Table 3:

Table 3. Primers used for qPCR

Primer	**sequence (5' →3')**
Gag forward	GGAGCTAGAACGATTCGCAGTTA
Gag reverse	GGTGTAGCTGTCCCAGTATTTGTC
WPRE forward	GGCACTGACAATTCCGTGGT
WPRE reverse	AGGGACGTAGCAGAAGGACG
Alb forward	GCTGTCATCTCTTGTGGGCTGT
Alb reverse	ACTCATGGGAGCTGCTGGTTC

3.10 Transduction of the leukemic blasts

Leukemic blasts (1 x 10^6 cells *per* well) are seeded in Retronectin-coated (Takara Bio) 24-well plates (40-60 μg/ml *per* well). The leukemic cells are transducted using a multiplicity of infection (MOI) of 2, in 1 ml of medium containing 8 mg/ml polybrene. Transduction is performed by spinoculation at 2300 rpm for 90 min at 32-37°C and incubation at 37°C for 2 hrs. Since the polybrene can be toxic for the cells, after this period of time, the infection medium is replaced with fresh medium and the cells are incubated for 24 hours at 37°C. After this incubation, half of the sample is used for transplantation into recipient mice and the other half is used to extract directly the DNA to be used as reference of infection at time of transplantation.

3.11 Measurment of the level of engrafment of hAML blasts by FACS analysis

The level of engrafment of the human leukemic cells in the transplanted animals is measured by FACS analysis of 20 µl Peripheral Blood cells stained with antibody specific for the human CD45 hematopoietic surface antigen (hCD45) APC conjugated (Miltenyi Biotec). For the staining, the plasma is removed by centrifugation at 2000 rpm. After blocking in 50 µL of 10% BSA in PBS 1X at 4°C for 30 min, the samples are resuspended in 50 µL di PBS 1X containing the specific antibody with a dilution 1:100. After 1-hour incubation at RT, the red blood cells are lysed in RBC Lysis Buffer (8.125mg/mL NH_4Cl, 1mg/mL $KHCO_3$, 0.13mM EDTA) for 10 min on ice. After 2 washing cycles, the cells are fixed in 1% FA (Formaldehyde, Sigma Aldrich) and analysed by FACS. The level of engraftment of leukemia in the host animals is determined as percentage of hCD45+ cells. The recipient mice are sacrificed when the percentage of hCD45+ is >60%. Bone marrow and spleen are collected and the purified blasts are frozen in 10% dimethyl sulfoxide (DMSO) in FBS and stored in liquid nitrogen.

3.12 Preparation of spike-in control cells

To prepare the spike-in controls, we selected 5 independent clones from a clonal tracking library containing 13,000 different barcodes (Cellecta). This library is similar to the 30 million clonal tracking library but contains a single barcode of 18 bp. Because the barcode is cloned in the same vector of the clonal tracking library, they can be amplified with the same PCR strategy described for the library in section 3.14. The list of the sequences of the 5 barcodes used as spike-in controls is given in Table 4.

Table 4. Barcodes sequences used for Spike in Experiment.

Clone	Barcode Sequence
Clone A	GTGTTGCATGCACAACTG
Clone B	ACTGTGGTTGTGACCACA
Clone C	TGGTTGCACATGCACACA
Clone D	CAACACGTACTGGTGTCA
Clone E	ACACACGTGTCATGTGAC

The 5 single clones were selected by bacterial transformation and isolation of single colonies. The bacterial colonies were grown for purification of the DNA by maxipreps. The DNA was used to make high-titer lentivirus (as described in section 3.8).

NB4 cells were infected at low MOI with each clone independently. We obtained a percentage of infection between 5 and 10%. The infected NB4 cells were, then, selected in 1μg/ml puromicyn for 48 hrs in order to obtain a pure population of cells with only one barcode *per* cell.

By serial dilution, we prepared 5 spike-in controls, that were combined with samples of blasts from which the barcode libraries were amplified for HTS:

- 10 NB4 cells with clone A
- 100 NB4 cells with clone B
- 1,000 NB4 cells with clone C
- 10,000 NB4 cells with clone D
- 50,000 NB4 cells with clone E

3.13 Genomic DNA extraction for amplification and sequencing of the barcodes

Identification of the barcodes in the experimental samples requires amplification of the integrated lentiviral constructs from the genomic DNA of the leukemic samples. Subsequent high-throughput sequencing of the amplified barcodes by the Illumina HiSeq 2000 platform is performed to identify and determine the relative frequency of each barcode in the tumor population.

The extraction of DNA from leukemic cells is performed by phenol-chloroform extraction according to the Cellecta Inc manual. Cells are resuspended at a concentration of 15 x 10^6 /mL in Resuspension Buffer (Macherey-Nagel), containing RNase A (400µg/mL) and 10% SDS and incubated for 5 min at RT. One volume of phenol:chloroform:isoamyl alcohol 25:24:1 (Sigma-Aldrich) is added, the mix is vigorously vortexed and centrifuged at 3250 x g for 90 minutes at RT. After centrifugation, the upper phase is collected and cleaned up from phenol residues with 1 volume of chloroform (VWR), vortexing and centrifuging at 3250 x g for 30 minutes at RT. The DNA is precipitated from the upper phase, adding 1 volume of isopropanol and 0.125 volumes of 3 M sodium. The mix is shacked vigorously and spinned down for 90 minutes a 3250 x g at RT. The pellet is washed with 70% ethanol by centrifugation at 3250 x g for 15 minutes at RT. The dried pellet is resuspended in sterile water and the concentration measured with the NanoDrop ND-1000 (Thermo).

3.14 Amplification of the barcodes from the genomic DNA

According to Cellecta Inc. suggestions, the library of barcodes from each sample is amplified starting from all the genomic DNA extracted from the leukemic blasts, because an adequate amount of DNA needs to be amplified in order to ensure full representation of the library of barcodes from all the cells isolated from each experimental sample. The lentiviral vectors of the libraries and the PCR primer are designed in order to include sequences complementary to the sequences of the immobilized primers necessary for generating amplification clusters in the Illumina's HiSeq 2000 flow cell. The amplification protocol consists of two round of PCR using a nested PCR strategy (Figure 10).

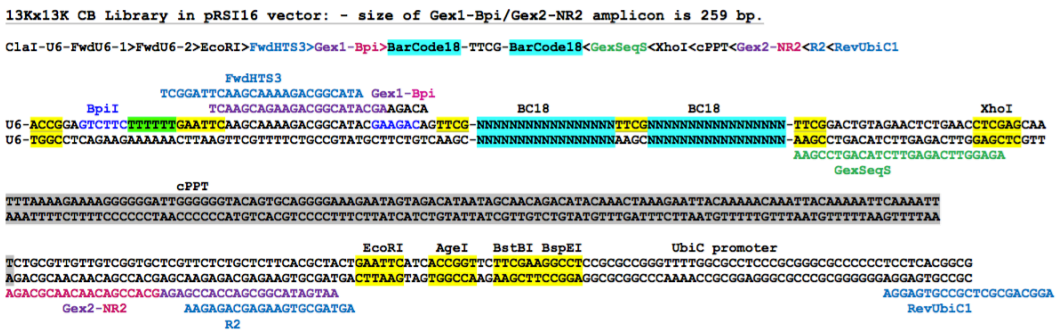

Figure 10. Cassette of the barcodes contained in the vectors of the library. The library is amplified using a nested PCR strategy that consists of two rounds of PCR. In the first round, primer FwdHTS3 and R2 are used to produce fragment of 275 bp. In the second round, the Gex-Bpi and Gex2-NR2 primers are used to amplify the 267 bp fragment ready to be sequenced on the Illumina platform.

The goal of the first PCR is to amplify barcodes from the genomic DNA while the goal of the second PCR, which uses only a sampling from the 1st PCR with nested primers, is to separate the amplified barcodes from non-specific PCR products and the excess of genomic DNA. All amplification are catalysed by the Titanium Taq DNA polymerase, very efficient in amplifications from very high amount of DNA (Clontech-Takara).

In detail, the PCR mix consists of: genomic DNA, 1X Titanium Taq Buffer, 0.5 U Titanium Taq polymerase, 400 µM of each dNTPs, 300 nM of each primer, H_2O up to 100 µl, with 25µg of genomic DNA *per* PCR sample. The PCR program is the following:

Temperature	Time	Cycles
94°C	3 min	1
94°C 66°C 72°C	30 sec 10 sec 20 sec	16
68°C	2 min	1

After the first round of PCR, for each sample all the independent PCR reactions are combined together in one aliquot and extensively mixed. 2µl of this mix are then used to perform the second round nested-PCR. The primers used for this second amplification, contain already the P5 and P7 adapter sequences for the Illumina's flow cell. The reaction mix is: 2µl of 1st PCR, 1X Titanium Taq Buffer, 0.5 U Titanium Taq Polymerase, 400 µM

each dNTPs, 500 nM each primer, H_2O up to 100 µl.

The PCR program is the following:

Temperature	Time	Cycles
94°C	3 min	1
94°C 66°C 72°C	30 sec 10 sec 10 sec	16/24
68°C	2 min	1

The specific primers for amplification of the library for clonal tracking and of the library for shRNA screening are listed in Table 5. 10 µl of the PCR products are analyzed by gel-electrophoresis in 1X TAE on a 3.5% agarose gel in order to confirm the correct size of the amplicons (267 bp both for Clonal Tracking and 259 for shRNA screening) and equal yields of the amplified barcodes for all samples. When necessary, in order to obtain the same yields among all amplified samples, the number of cycles are adjusted up to a maximum of 24 cycles. The correct size PCR products are purified using QIAquick Gel Extraction (QIAGEN) commercial kit. The quantification of the amplicons is determined by Qubit Assay (Invitrogen).

Since the lentiviral shRNA library and PCR primer were designed to obtain sequences complementary to the sequences of the immobilized primers P5 and P7 necessary for generating amplification clusters in Illumina's HiSeq 2000, the amplicons were ready to be sequenced through GexSeq primer.

Table 5. Primer used for amplification of the barcodes library from the leukemic blasts. As already described, the P5 and P7 sequences (blue and red, respectively) are contained in the 2nd round primers.

Experiment	Round	Name	Sequence
Clonal Tracking	1st	FwdHTS3	TCGGATTCAAGCAAAAGACGGCATA
		R2	AGTAGCGTGAAGAGCAGAGAA
	2nd	Gex1-Bpi	TCAAGCAGAAGACGGCATACGAAGACA
		NR2	AATGATACGGCGACCACCGAGACGAGCACCGACAACAACGCAGA
shRNA screening	1st	FwdHTS	TCGGATTCGCACCAGCACGCTA
		R2	AGTAGCGTGAAGAGCAGAGAA
	2nd	Gex1-NF2	CAAGCAGAAGACGGCATACGATTCGCACCAGCACGCTACGCA
		Gex2-NR2	AATGATACGGCGACCACCGAGACGAGCACCGACAACAACGCAGA
Sequencing		GexSeqS	AGAGGTTCAGAGTTCTACAGTCCGAA

3.15 Cloning of shRNAs in pInducer11 inducible vector

For validation of the target genes identified by our shRNA screening, we cloned the corresponding shRNAs in the pInducer11 vector. For each shRNA two different copies of complementary oligonucleotides (a top and bottom strand) were ordered from SIGMA Aldrich. The sequences were the same of the shRNAs present in the library from Cellecta. All the forward sequences, top strand, had the same basic design, starting from 5': XhoI restriction site (TCGAG), a conserved sequence (ACCGG), 21 nucleotides (specific for each shRNA) target sense sequence with mismatches, hairpin loop (GTTAATATTCATAGC), 21 nucleotides target antisense sequence (specific for each shRNA), terminator sequence of 4 thymines and a G to complete EcoRI restriction site. These sequences allow the unilateral cloning through the complementarity of restriction site. The complete sequences, both for the forward and reverse strands, are listed in Table 6.

The forward and reverse complementary oligonucleotides were mixed in a 1:1 ratio at 2µM each. The annealing reaction was performed in 50 µl of Annealing buffer 1X (100 mM CH_3COOK, 30 mM HEPES-KOH, 2 mM Mg-acetate), incubating at 95°C for 4 minutes, 72°C for 10 min, and at room temperature (RT) over night (O/N). The annealed oligonucleotides were stored at -20°C until use.

The annealed oligonucleotides were phosphorylated in order to increase the cloning efficiency. The reaction was performed incubating at 37°C for 30 min in a 50 µl volume with 5 µl of 10x T4 DNA ligase buffer (that contains 1 mM ATP), 1 µl (10 units) of T4 Polynucleotide Kinase (T4 PNK) (Neb) and 25 µl (1 µM) of annealing reaction. The reaction was stopped by heat-shock inactivation at 65°C for 20 min.

To clone the phosphorylated oligonucleotides, the pInducer11 plasmid was digested with XhoI (10 Units/µg) and EcoRI (10 Units/µg) for 1 hour at 37°C. The reaction was performed in 1x CutSmart® Buffer (Neb). After 1 hour, the digested linearized vector was

separated by electrophoresis on a 0.8% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen), according to the manufacturer's specifications.

2 nM of the phosphorylated oligonucleotides and 75 ng of linearized vector were ligated using 1 µl (10 Units) of T4 DNA Ligase (NEB). After transformation of the ligation mix in TOP10 *E.*coli, positive clones were screened by Colony PCR.

Table 6. List of the oligonucleotides cloned in the pInducer vector for validation. The targets were identified by our shRNA screening. In capital letters are indicated the shRNA sequences; in lowercase letters the sites for restriction enzyme XhoI (tcgag) and EcoRI (aattc). Name Oligo, name of the target gene and number of the shRNA in the Cellecta library; f, forward; r, reverse.

Name Oligo	Sequence (5'à3')
BRD9.743f	tcgagACCGGCAATGAAGATACAGTTGTTGAGTTAATATTCATAGCTCAACAGCTGTATCTTCATTGTTTTg
BRD9.743r	aattcAAAACAATGAAGATACAGCTGTTGAGCTATGAATATTAACTCAACAACTGTATCTTCATTGCCGGTc
BRD9.747f	tcgagACCGGGAGAGCACACTTATTCAGTAAGTTAATATTCATAGCTTGCTGAATAGGTGTGCTCTCTTTTg
BRD9.747r	aattcAAAAGAGAGCACACCTATTCAGCAAGCTATGAATATTAACTTACTGAATAAGTGTGCTCTCCCGGTc
CBX5.593f	tcgagACCGGGCCGATGACATCAAATTTAAAGTTAATATTCATAGCTTTAGATTTGATGTCATCGGCTTTTg
CBX5.593r	aattcAAAAGCCGATGACATCAAATCTAAAGCTATGAATATTAACTTTAAATTTGATGTCATCGGCCCGGTc
CBX5.977f	tcgagACCGGGAGAGAGCAGAGTAATGATATGTTAATATTCATAGCATATCATTGCTCTGCTCTCTCTTTTg
CBX5.977r	aattcAAAAGAGAGAGCAGAGCAATGATATGCTATGAATATTAACATATCATTACTCTGCTCTCTCCCGGTc
JMJD6.605f	tcgagACCGGGCGGTATGAAAGACCTTATAAGTTAATATTCATAGCTTGTAAGGTCTTTCATACCGCTTTTg
JMJD6.605r	aattcAAAAGCGGTATGAAAGACCTTACAAGCTATGAATATTAACTTATAAGGTCTTTCATACCGCCCGGTc
JMJD6.972f	tcgagACCGGATGGACTTTGGAGCGCTTAAAGTTAATATTCATAGCTTTAGGCGCTCCAGAGTCCATTTTTg
JMJD6.972r	aattcAAAAATGGACTCTGGAGCGCCTAAAGCTATGAATATTAACTTTAAGCGCTCCAAAGTCCATCCGGTc
KDM4C.449f	tcgagACCGGGCCCAGGTCTTGGTATGTTATGTTAATATTCATAGCATAGCATACCAAGACTTGGGCTTTTg
KDM4C.449r	aattcAAAAGCCCAAGTCTTGGTATGCTATGCTATGAATATTAACATAACATACCAAGACCTGGGCCCGGTc
KDM4C.920f	tcgagACCGGGCTATGAGAAGCCTGAGAAGTGTTAATATTCATAGCATTTCTCGGGCTTCTCATAGCTTTTg
KDM4C.920r	aattcAAAAGCTATGAGAAGCCCGAGAAATGCTATGAATATTAACACTTCTCAGGCTTCTCATAGCCCGGTc
KDM5D.455f	tcgagACCGGGCCACATTGGAAGCTATAATTGTTAATATTCATAGCAATTATGGCTTCCAATGTGGCTTTTg
KDM5D.455r	aattcAAAAGCCACATTGGAAGCCATAATTGCTATGAATATTAACAATTATAGCTTCCAATGTGGCCCGGTc
KDM5D.869f	tcgagACCGGTGTATCTTGGTGGAGTGTAAAGTTAATATTCATAGCTTTACACTCCGCCAAGATACATTTTg
KDM5D.869r	aattcAAAATGTATCTTGGCGGAGTGTAAAGCTATGAATATTAACTTTACACTCCACCAAGATACACCGGTc
Luc.01f	tcgagACCGGCTTCGAAATGTTCGTTTGGTTGTTAATATTCATAGCAACCGAACGGACATTTCGAAGTTTTg
Luc.01r	aattcAAAACTTCGAAATGTCCGTTCGGTTGCTATGAATATTAACAACCAAACGAACATTTCGAAGCCGGTc
Luc.02f	tcgagACCGGTCACAGAATCGTTGTATGTAGGTTAATATTCATAGCCTGCATACGACGATTCTGTGATTTTg
Luc.02f	aattcAAAATCACAGAATCGTCGTATGCAGGCTATGAATATTAACCTACATACAACGATTCTGTGACCGGTc
PRDM6.911f	tcgagACCGGGTTCAGTACAGGTTGAATATAGTTAATATTCATAGCTATATTCGACCTGTACTGAACTTTTg
PRDM6.911r	aattcAAAAGTTCAGTACAGGTCGAATATAGCTATGAATATTAACTATATTCAACCTGTACTGAACCCGGTc
PRDM6.915f	tcgagACCGGCTACCGAGCTTGTATAGATATGTTAATATTCATAGCATATCTATACAGGCTCGGTAGTTTTg
PRDM6.915r	aattcAAAACTACCGAGCCTGTATAGATATGCTATGAATATTAACATATCTATACAAGCTCGGTAGCCGGTc
PRDM9.224f	tcgagACCGGGCAGCAATATCTAGATCCATAGTTAATATTCATAGCTGTGGATCTGGATATTGCTGCTTTTg
PRDM9.224r	aattcAAAAGCAGCAATATCCAGATCCACAGCTATGAATATTAACTATGGATCTAGATATTGCTGCCCGGTc
PRDM9.990f	tcgagACCGGTTAAGAGTGGAACAGTGTAAAGTTAATATTCATAGCTTTACGCTGTTCCACTCTTAATTTTg
PRDM9.990r	aattcAAAATTAAGAGTGGAACAGCGTAAAGCTATGAATATTAACTTTACACTGTTCCACTCTTAACCGGTc
PRMT1.500f	tcgagACCGGGTGTTCCAGTATCTTTGATTAGTTAATATTCATAGCTAATCAGAGATACTGGAACACTTTTg
PRMT1.500r	aattcAAAAGTGTTCCAGTATCTCTGATTAGCTATGAATATTAACTAATCAAAGATACTGGAACACCCGGTc
PRMT1.951f	tcgagACCGGCTTACCGTAACTCTATGTTTCGTTAATATTCATAGCGAAACATGGAGTTGCGGTAAGTTTTg
PRMT1.951r	aattcAAAACTTACCGCAACTCCATGTTTCGCTATGAATATTAACGAAACATAGAGTTACGGTAAGCCGGTc
Psma1.011f	tcgagACCGGGCAATGGAAGCTGTTAAATAAGTTAATATTCATAGCTTGTTTAACAGCTTCCATTGCTTTTg
Psma1.011r	aattcAAAAGCAATGGAAGCTGTTAAACAAGCTATGAATATTAACTTATTTAACAGCTTCCATTGCCCGGTc
Psma1.020f	tcgagACCGGGAAAGGGTCTGTATAATCATTGTTAATATTCATAGCAATGATTATACAGACCCTTTCTTTTg
Psma1.020r	aattcAAAAGAAAGGGTCTGTATAATCATTGCTATGAATATTAACAATGATTATACAGACCCTTTCCCGGTc
Rpl30.001f	tcgagACCGGCCATCACTACAGTGGTAATAAGTTAATATTCATAGCTTATTGCCACTGTAGTGATGGTTTTg
Rpl30.001r	aattcAAAACCATCACTACAGTGGCAATAAGCTATGAATATTAACTTATTACCACTGTAGTGATGGCCGGTc
Rpl30.009f	tcgagACCGGGCTTTGAGGAAATCTGAAATAGTTAATATTCATAGCTATTTCAGATTTCCTCAAAGCTTTTg
Rpl30.009r	aattcAAAAGCTTTGAGGAAATCTGAAATAGCTATGAATATTAACTATTTCAGATTTCCTCAAAGCCCGGTc
SETDB1.828f	tcgagACCGGCAGTGATTAATTGTGAGTCTTGCTATGAATATTAACAAGACTCACAATTAGTCACTGTTTTg
SETDB1.828r	aattcAAAACAGTGACTAATTGTGAGTCTTGCTATGAATATTAACAAGACTCACAATTAATCACTGCCGGTc
SETDB1.930f	tcgagACCGGAGTTAGAGACATGGGTAATATGTTAATATTCATAGCGTATTACCCATGTCTCTAACTTTTTg
SETDB1.930r	aattcAAAAAGTTAGAGACATGGGTAATACGCTATGAATATTAACATATTACCCATGTCTCTAACTCCGGTc
TRIM24.395f	tcgagACCGGCGCATGAAACTTATGTAACAAGTTAATATTCATAGCTTGTTGCATAAGTTTCATGCGTTTTg
TRIM24.395r	aattcAAAACGCATGAAACTTATGCAACAAGCTATGAATATTAACTTGTTACATAAGTTTCATGCGCCGGTc
TRIM24.399f	tcgagACCGGCGAGACTTATCTAAACTAGAAGTTAATATTCATAGCTTCTGGTTTAGATAAGTCTCGTTTTg
TRIM24.399r	aattcAAAACGAGACTTATCTAAACCAGAAGCTATGAATATTAACTTCTAGTTTAGATAAGTCTCGCCGGTc
TRIM66.893f	tcgagACCGGCTGTCATTGCATGGTTAATTTGTTAATATTCATAGCAAATTAGCCATGCAATGACAGTTTTg
TRIM66.893r	aattcAAAACTGTCATTGCATGGCTAATTTGCTATGAATATTAACAAATTAACCATGCAATGACAGCCGGTc
TRIM66.896f	tcgagACCGGCCGGAGATTTCACTTTGTATTGTTAATATTCATAGCAATACAAGGTGAAGTCTCCGGTTTTg
TRIM66.896r	aattcAAAACCGGAGACTTCACCTTGTATTGCTATGAATATTAACAATACAAAGTGAAATCTCCGGCCGGTc

3.16 Transformation of DNA into competent *E. coli* TOP10 cells

After ligation, 25ng of DNA were used to transform *E. coli* TOP10 competent cells. Briefly, *E. coli* TOP10 competent cells were thawed on ice for 10 minutes. 25 ng of plasmid DNA was added to the cells and the tubes were carefully flicked to mix cells and DNA. The mixtures were:

- placed on ice for 30 minutes.
- heat-shocked at exactly 42°C for 60 seconds
- placed on ice for 5 minutes without mixing
- placed at 37°C for 60 minutes after addition of 250 µl of LB. During this phase the cells were shacked vigorously (250 rpm).
- plated (50-100 µl of each transformation) onto a selection plate (50 mg/L ampicillin) and incubated overnight at 37°C.

3.17 Colony PCR for screening of positive clones

After transformation in *E.coli* TOP10, positive clones were screened by Colony PCR. Individual transformants were added directly to the PCR reaction and lysed during the initial heating step, which causes the release of the plasmid DNA from the bacteria. The primers for PCR reaction (Table 7) were designed to specifically amplify a band of 448 bp if the construct contains the shRNA insert. For this purpose, the colonies to be tested were picked with a sterile tip and, after replating, directly inoculated in the appropriate PCR mix (1X Taq Polymerase Buffer, 400 mM forward and reverse primers, 1U Taq Polimerase (NEB), 200 mM dNTPS) and amplified according to the following protocol:

Temperature	Time	Runs
95°C	2 min	1
95°C 47°C 72°C	1 min 1 min 2 min	30

Table 7. Primers for colony PCR.

Primer name	Sequence (5'→3')
pINDfc	CACATACAGATCCAAGAAACC
pINDr	GTATCAAAGAGATAGCAACG

Positive clones were inoculated in an appropriate volume of LB, containing ampicillin (100 mg/L), and grown at 37°C, shacking vigorously (250 rpm) O/N. For DNA extraction NucleoBond Xtra Maxi or NucleoSpin Plasmid (noLid) (Macherey-Nagel) were used according to the volume of growth and the manufacturer's instructions. The purified plasmid DNA was quantified using NanoDrop ND-1000 (Thermo). Quality assessmant was based on the ratios of absorbance 260nm/280nm and 260nm/230nm. For DNA a 260/280 ratio of ~1.8 and 260/230 of 2.0-2.2 was generally accepted as "pure".

4 RESULTS

4.1 Establishment of a bank of AMLs

Instrumental for this study was the collection of both murine and human leukemic samples, available for our experiments without limits of accessibility. For this reason we firts established a bank of leukemic samples. In particular, since human acute myeloid leukemia (hAML) xenotransplanted in immunocompromised animals retain phenotypic characteristics of the original sample, we established xenografts of blasts collected from patients affected by AMLs. As host, we used immunodeficient mice with components of the human immune system, the NOD SCID IL2RG null (NSG) immunocompromised mice. NSG mice have been shown to support greater engraftment rates of human hematopoietic stem cells (hCD34+ cells) than all other strains[75]. Our protocol, as described in details in the Materials and Methods, consists of the intravenous injection of CD3-depleted mononuclear bone marrow (BM) or peripheral blood cells freshly collected from AML patients in 6-8 weeks old NSG mice. The development of the disease is monitored by the level of engrafment of the human blasts in the peripheral blood (PB) of the transplanted animals by FACS analysis, after staininig of the PB cells with antibody specific for the human CD45 hematopoietic surface antigen (hCD45). Up to date, we transplanted AML blasts collected from 77 patients and obtained successful engrafment of 33, of which 16 could be serially passaged in secondary recipients in order to expand, collect and store cells for subsequent experiments. All the successfully engrafted leukemias have been characterized for their immunophenotype and main cytogenetic alterations. Moreover, the samples have been sequenced by Sanger for detection of mutations in NPM1 and FLT3, the most frequently mutated genes in AMLs. In our lab we have now established a bank of xenotransplanted hAMLs representative of leukemias with diverse clinical features belonging to different risk stratification groups.

The samples that we were able to infect in our study and their main genetic characteristics are reported in Table 8.

Table 8. Xenotransplanted human AMLs used in our study: leukemic samples were characterized for the principal recurrent mutations, NPM1c+ and FLT3-ITD, and for their karyotype. WT: Wild-type; MUT: mutated.

Sample ID	NPM1	FLT3	kariotype
AML-IEO20	WT	WT	46, XX; t(9;11)
AML-IEO23	MUT	MUT	Normal
AML5	WT	MUT	46, XY; t(1;2)
AML9	WT	WT	Complex

For the murine AMLs, we collected several samples from transgenic animal models characterized by the expression of the fusion protein PML-RAR (mAPL), the mutant NMP1 (NMPc+) and the combination of NPMc+ with Flt3-ITD (NMP/Flt3). We started our experiments on the APL model, that will be presented in this thesis.

The mCG-PML-RARalfa knock in (KI) mice[43] express the PML-RAR fusion protein under the control of the catepsin-G promoter. These animals develop spontaneous leukemias with a frequency of ~55% and an average latency of 11 months. The developed leukemias are clinically and phenotypically undistinguishable from the promyelocytic leukemias developed in human patients that harbor the translocation t(15;17). We collected blasts from the spleen and bone marrow of PML-RAR KI mice at signs of disease, expanded them in secondary recipients and froze the cells in several aliquots for our experiments.

4.2 Experimental approach

Our experimental approach has a common workflow for both the clonal tracking analysis and the *in vivo* screenings of shRNA libraries (Figure 11). It is based on the infection of freshly-isolated leukemic blasts with lentiviral libraries and their transplantation in recipient animals. We infect blasts at low multiplicity of infection (MOI), to ensure single retroviral integration *per* cell. After infection, we divide the leukemic blasts in two pools. From the first pool we extract genomic DNA to be used as the reference starting point, after infection *in vitro*. The second pool is transplanted in recipient animals. At the onset of leukemia in the transplanted animals, usually when the engraftment reaches about the 90%,

we collect leukemic blasts and extract genomic DNA. From genomic DNA of the two pools we prepare libraries of barcodes by PCR amplification and submitt to high throughput sequencing by Illumina. Bioinformatic analysis of the barcodes allows comparison of library composition before and after tumor growth *in vivo* (Figure 11). These analyses allow us to identify and track the fate of the different LICs for the clonal tracking experiments and to identify biologically relevant genes in the shRNA screening, as described in more details in the following sections.

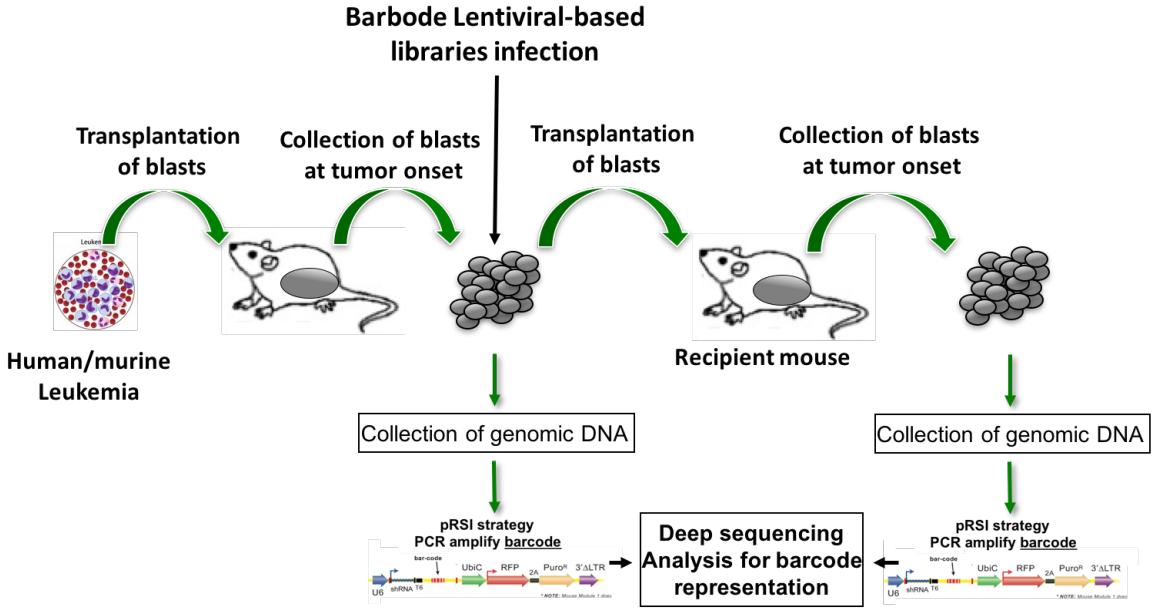

Figure 11. Workflow of our experimental approach. Freshly isolated leukemic cells, derived from mouse models or from xenotransplanted human samples, are infected at low MOI with our lentiviral libraries. After infection, half of the sample is used as a reference of infection *in vitro* and the remaining half is transplanted into recipient mice. After development of the leukemia *in vivo*, the recipients are sacrificed and the main organs and tissues (spleen and bone marrow) infiltrated by the leukemic cells are collected. From both samples, we extract the genomic DNA and the libraries of barcodes are prepared by PCR amplification and sent to sequencing by Illumina. Analysis of the barcodes is performed using bioinformatics pipelines in order to compare the composition of barcodes in the two libraries, right post-infection and following leukemia onset.In the clonal tracking experiments, after development of the leukemia in the primary recipient, part of the collected blasts are serially transplanted in secondary and tertiary recipients in order to follow the evolution of the disease.

We first defined appropriate experimental conditions for our *in vivo* studies: infectability of the primary leukemic blasts with the lentiviral vectors of our libraries and their transplantability in recipient animals. We tested different protocols of infection and found

that spinoculation at 2300 rpm for 90 min at RT as the most efficient. Moreover, since primary leukemic blasts tend to lose viability in culture, we infected the blasts immediately after isolation from transplanted leukemic mice. Even using the same protocols and MOI, we observed that the efficiency of infection, measured as percentage of TagRFP-positive (RFP+) cells by FACS analysis, is very variable among samples and some leukemias failed to be infected. However, once infected, the leukemic blasts, after transplantation in recipient animals, are able to engraft and tend to maintain the same percentage of RFP+ cells following leukemia development *in vivo* (Figure 13), suggesting that the lentiviral infection does not affect homing and growth properties of LICs. Notably, the infected blasts are not counterselected during *in vivo* leukemogenesis.

A critical aspect of both clonal tracking and shRNA screenings, is the presence of single integrants and, as a consequence, single barcodes *per* cell. In the clonal tracking, in fact, in order to be able to track single cells independently, each infected cell must be identified by a unique barcode. In the shRNA screening, instead, we need a single integrant because we need to ensure that the observed phenotype is supported by knock down of a single gene *per* blast. To this aim, we evaluated the number of integrants *per* cell by qPCR according to the protocol described in Material and Methods. We determined that an efficacy of infection of around 20% assures a single integrant *per* cell. Because, in our experimental conditions, an efficiency of infection <20% is usually obtained with a MOI of 2, we infect the leukemic blasts with this MOI. As reported in section 4.4.1 this MOI guarantees us to have one integration *per* cell.

4.3 Clonal tracking experiments

The rationale of our clonal tracking experiments is to mark each leukemic cell with a specific genetic marker, the barcode, that can be stably transmitted to its progeny and tracked in a complex cellular population to establish their contribution to leukemia development *in vivo*. Thus, in our clonal tracking analyses we considered two parameters:

the identity of the barcode, that identify each leukemic clone, and the frequency of each barcode in the leukemic population, as indirect measurement of the contribution of each clone to leukemia development. In more details, for each sample, the sequences (reads) derived from the Illumina platform are aligned to the reference library provided by the manufacturer (Cellecta Inc.), without allowing mismatches. For each barcode identified, we calculate the relative frequency in the tumor sample, as the ratio between the reads that identify that specific barcode and the total number of reads obtained from the sequencing of the library for each sample.

4.3.1 Infection of AML blasts with the clonal tracking library

To track clones originated by individual leukemic cells inside the tumor population, we infect leukemic blasts with a lentiviral library containing 30 million different barcodes. The complexity of the library is generated by 30 million different combinations of two barcodes of 18 degenerated nucleotides each. This very high complexity allows to mark each cell with a unique barcode. Indeed, infecting a low number of cells (1 million) with a library that contains a very high number of different barcodes minimize the probability of obtaining integration of the same barcode in two different cells. In particular, for our studies we infected human blasts of one xenotransplanted human AML, AML-IEO20, and murine blasts from a transgenic PML-RAR KI mouse, mAPL18. In order to perform our clonal tracking analysis in parallel in different animals, for both leukemias, we infected the blasts with the 30 million-barcodes library in 5 independent cell aliquots and transplanted each of them in 5 recipients animals (Figure 12). We determined by FACS analysis the percentage of cells expressing the reporter gene TagRFP (RFP) and obtained ~15% infection for AML-IEO20 and ~8% for mAPL18. For two of these recipients, chosen on the basis of TagRFP signal in primary mice, we serially transplanted the leukemic blasts up to the third passage (X3) for AML-IEO20 and up to the second passage (X2) for mAPL18 (Figure 12). 5x105 blasts were retransplanted in each secondary

recipient in serial passages. From each recipient, at the onset of leukemia, we collected the blasts, extracted the genomic DNA, amplified by PCR the barcodes and sequenced them by NGS.

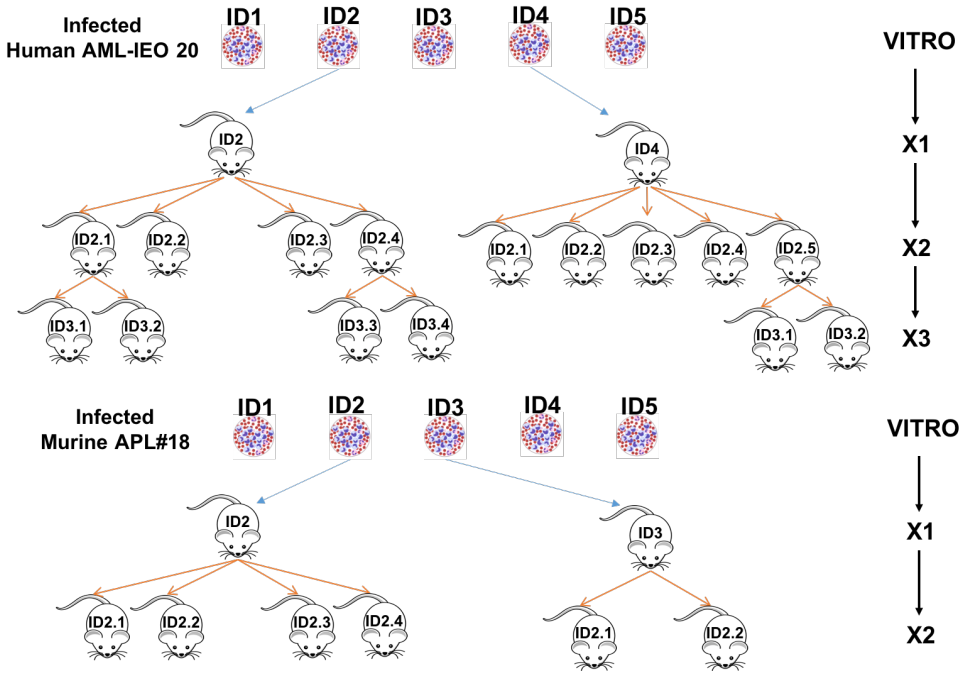

Figure 12. Scheme of the clonal tracking experiments. For each leukemia, AML-IEO20 or mAPL18, we infected with the 30 million barcodes library 5 independent pools of blasts (infection). For each leukemia five pools were transplanted in independent primary recipient mice (X1). The blasts collected from the primary leukemias, indicated with the arrows, were transplanted in several secondary recipients (from 2 to 5 mice, X2) and, in particular for AML-IEO20, the blasts collected from the secondary leukemias, were transplanted in tertiary recipients (X3). Each recipient mouse is identified by an ID number (ID).

For AML-IEO20, following leukemia development *in vivo*, all transplanted animals maintained, in all the hematopoietic compartments analysed (peripheral blood, spleen and bone marrow), a percentage of RFP+ cells similar to the one observed following infection *in vitro* (Figure 13A and 13B). However, while following the primary transplantation the percentage of RFP+ cells was very similar in all 5 recipient animals (Figure 13A and 13B), in the secondary recipients the percentage of RFP expression differed in recipients of leukemic blasts deriving from different primary donors (Figure 13C). Nonetheless, the RFP signal was very consistent for all the recipients of the same donor (Figure 13C).

Taken together, these results indicate that the infected blasts are not counterselected *in vivo* and the leukemogenic potential of LICs is not affected by infection with the clonal tracking library.

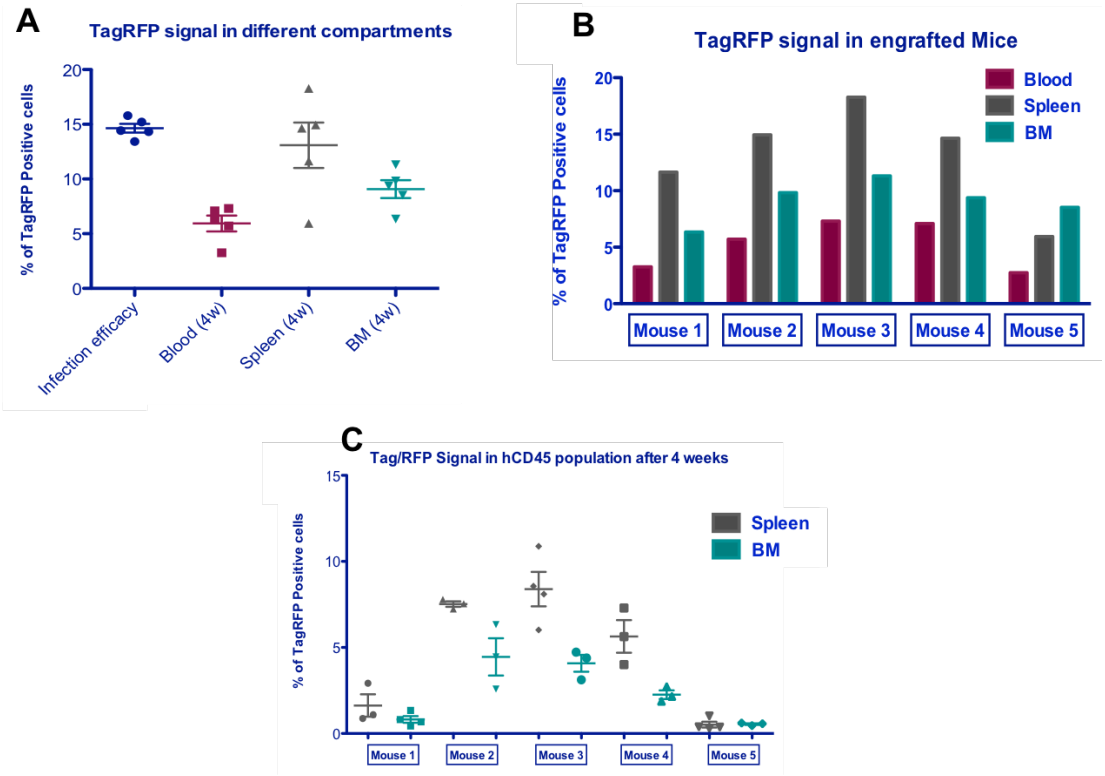

Figure 13. Efficacy of infection of AML-IEO20 and engrafment of the infected blasts following transplantation *in vivo*. A) Leukemic blasts from AML-IEO20 were infected with the 30 million barcodes library in 5 sets of 1 million cells each. After infection, half of the 5 sets of cells were transplanted independently in 5 recipient mice. At the onset of leukemia, the leukemic blasts were collected from different hematopoietic compartments: peripheral blood (Blood), spleen (Spleen) and bone marrow (BM). The level of RFP signal (% TagRFP positive cells) was evaluated by FACS in each compartment for each recipient. Each symbol represents a recipient animal. B) Panel B shows the level of RFP signal in the 3 leukemic compartments analysed for each recipient animal (Mouse). C) Level of engrafment of AML-IEO20 in secondary recipients. At the onset of leukemia, the leukemic cells were collected from spleen (Spleen) or bone marrow (BM) of each primary recipient (Mouse) and were retransplanted in 3 or 4 secondary recipients. As described for the primary leukemias, the RFP signal was evaluated in the peripheral blood of every secondary recipient at the onset of the tumors. Labels are as for panel A.

From all pools of leukemic cells collected after development of the leukemia *in vivo* the barcodes libraries were amplified by PCR from the genomic DNA of the collected blasts. As an example, Figure 14 shows the amplified library for the primary tumors. For each

tumor, after PCR amplification, we obtained a specific band corresponding to the expected size of 267 bps of the barcodes.

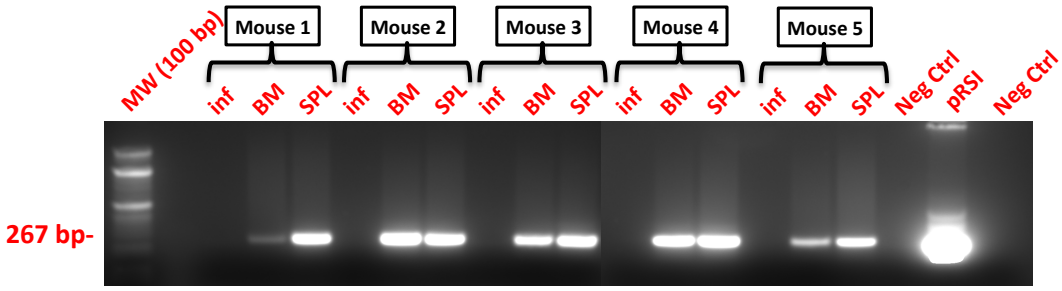

Figure 14. Amplification of the libraries of barcodes from AML-IEO20. The libraries were prepared by PCR amplification from the genomic DNA extracted from blasts of AML-IEO20 infected with the 30 million barcodes. Inf, blasts after infection *in vitro*. BM (bone marrow) and SPL (spleen) from animals transplanted with the infected primary leukemias. The PCR products were analysed on a 3.5% agarose gel and appeared to be specific and of the expected size (267 bps). MW, 100 bp Marker; pRSI, purified vector used as positive control; Neg Ctrl, Genomic DNA used as negative controls.

In contrast to what we observed for the human AML, for the murine mAPL18 leukemia, at the onset of the disease after transplantation, all the primary recipients showed a very low percentage (~1%) of RFP+ cells (Figure 15A). This level of RFP+ cells was consistent in all hematopoietic compartments analyzed (Figure 15A and 15B). Loss of RFP expression can be explained by two hypothesis: i) the RFP+ positive cells were counterselected *in vivo* in the mAPL18; ii) PML-RAR downregulates the expression of the reporter gene by gene silencing, as it has been previously reported[97]. In order to test the first hypothesis and determine if the infected blasts are indeed ablated by the leukemic population during *in vivo* growth, we tested by PCR amplification if the lentiviral integrations were still detectable in the leukemic blasts collected from the primary recipients animals. We performed a PCR on the genomic DNA of these blasts using primers specific for the lentiviral sequences GAG and WPRE. As shown in Figure 15C, with both copies of primers, in 4 out of the 5 transplanted animals we amplified specific bands, corresponding

to the viral integrations. These data suggest that the infected cells are not counterselected and more likely there is silencing of the reporter gene.

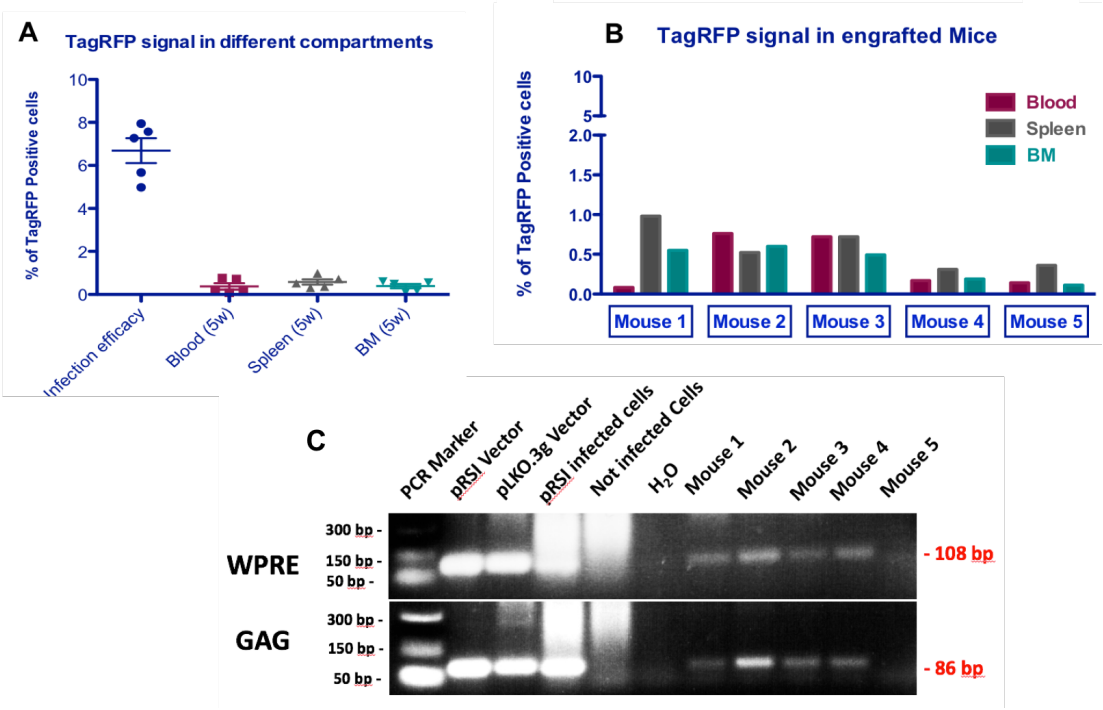

Figure 15. Efficacy of infection of mAPL18 and engrafment of the infected blasts following transplantation *in vivo*. Leukemic blasts from mAPL18 were infected with the 30 million-barcodes library in 5 sets of 1 million cells each. After infection, half of the 5 sets of cells were transplanted independently in 5 recipient mice. At the onset of leukemia, the leukemic blasts were collected from different hematopoietic compartments: peripheral blood (blood), spleen (spleen) and bone marrow (BM). Panel A shows the RFP signal in all compartments analyzed at the onset of leukemia compared to the signal detected in the blasts right after infection and prior to transplantation (Infection). For each compartment, each symbol represents a recipient animal. B) The RFP signal (% TagRFP positive cells) was evaluated in each compartment for each recipient (Mouse) when leukemic engrafment in the PB reached ~90%. C) Because we observed loss of RFP signal following leukemia onset, we extracted the genomic DNA from the blasts and determined by PCR amplification if the viral integrations were still detectable. We used two sets of primers specific for lentiviral sequences: GAG and WPRE. Specific bands were detectable in 4 out of 5 transplanted animals. As positive controls we used: two lentiviral vectors carrying GAG and WPRE sequences (pRSI and pLKO.3g vectors) and murine genomic DNA from a murine leukemia infected at high MOI (pRSI infected cells). As negative controls we used: genomic DNA from not infected cells (Genomic DNA) and no DNA (H_2O).

Indeed, in our laboratory, it as been previously observed that in this same leukemic model, administration of retinoic acid (RA) *in vivo* induces the rescue of the signal of the reporter gene GFP, introduced in the leukemic blasts by lentiviral infection. In that particular

experiment, we infected the blasts of mAPL18 with a lentiviral vector encoding for a fusion gene between histone H2B and GFP (H2B-GFP). We obtained an efficacy of infection of ~40% (Figure 16). However, at the onset of the leukemia, we could no detect any H2B-GFP+ blast in the peripheral blood of any of the recipient mice. Suspecting a silencing of the reporter due to the presence of PML-RAR (Figures 15A and 16). We, therefore, treated the leukemic mice with RA (40mg/Kg every other day for 5 administrations). As shown in Figure 16, after RA administration, we observed reexpression of the H2B-GFP in up to ~60% of the circulating blasts, confirming our hypothesis that PML-RAR induced silencing of reporter gene.

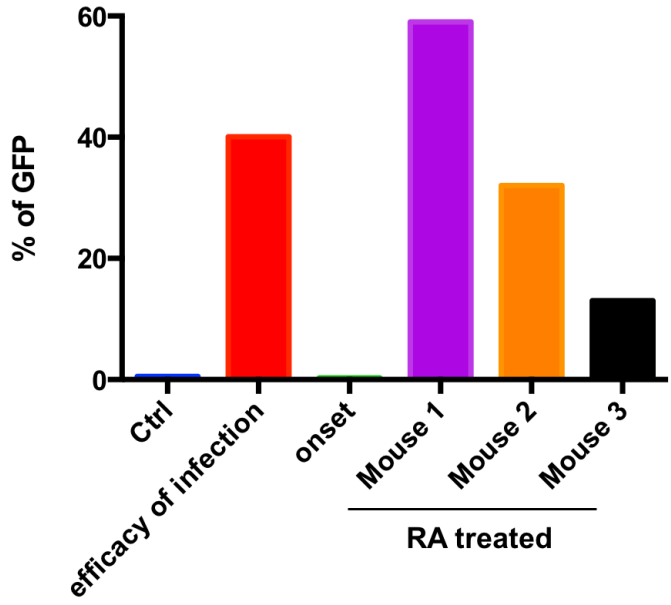

Figure 16. Retinoic acid treatment induces reexpression of the H2B-GFP fusion protein in the mAPL18 blasts. The percentage of blasts expressing the H2B-GFP (% of GFP) was evaluated by FACS analysis in the leukemic blasts: i) after infection *in vitro* (efficacy of infection); ii) at the onset of the leukemia in one representative transplanted recipient (onset); iii) in the transplanted recipients (Mouse 1, Mouse 2 and Mouse 3) after treatment with 40 mg/kg of reinoic acid (RA treated). Ctrl, negative control: not transplanted mouse.

We are currently trying to address this issue in the blasts infected with the clonal tracking library. We already transplanted new recipient mice with the blasts derived from one of

these mAPL18 samples and we plan to treat them with pharmacological doses of retinoic acid in order to see if we are able to resume RFP expression.

Considering that the infected cells were still detectable in the leukemic samples, we decided to proceed with the amplification of the library of barcodes from the genomic DNA extracted from all pools of leukemic cells. In order to established a threshold for clonal identification, before amplification by PCR of the library of barcodes, we added a spike-in control to the genomic DNA of each sample. As spike-in control, we used genomic DNA extracted from NB4 cells infected with specific barcodes (described in detail in the following paragraph 4.3.2). Figure 17 shows an example of the PCR amplification for the libraries from the primary tumors of mAPL18. For each tumor, we obtained two specific bands: one corresponding to the expected size of 267 bps for the barcodes of clonal tracking library and one of 249 bps for the spike-in control barcodes (Figure 17).

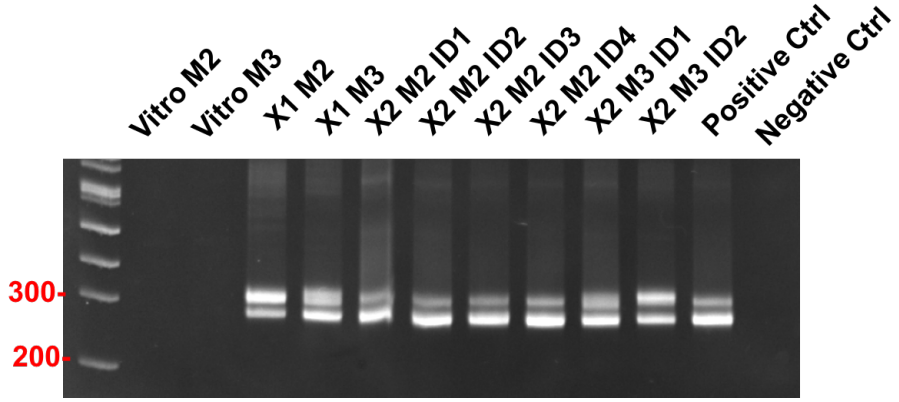

Figure 17. Amplification of the libraries of barcodes from mAPL18. The libraries were prepared by PCR amplification from the genomic DNA extracted from blasts of mAPL18 infected with the 30 million barcodes. The PCR products were analysed on a 3.5% agarose gel and appeared to be specific and of the expected size (267 bps for the barcodes of clonal tracking library and one of 249 bps for the spike-in control barcodes). Vitro M2 and M3, blasts after infection *in vitro*. MW, 100 bp Marker; Positive Ctrl, purified vectors used as positive control; Negative Ctrl, Genomic DNA of not infected cells used as negative control.

For both the human and the murine leukemia, for all samples with successful amplification, after gel purification, the barcode-libraries were sequenced with the Illumina HiSeq platform.

4.3.2 Establishment of a threshold for clone identification

One major issue in this type of experiments is the establishement of a proper threshold for the identification of the sequence tags to be tracked during the clonal analysis. In order to perform internal calibration and correct for technical issues related to genomic DNA extraction and to amplification by PCR, we added an internal control to each experimental sample of the murine leukemia, during the phase of genomic DNA extraction and preparation of the libraries for sequencing. These ‘‘spiked-in’’ controls consisted of defined numbers of clonal NB4 cells (10, 100, 1,000, 10,000 and 50,000 cells) infected with lentiviral constructs each harboring a known single barcode of 18 nucleotides, that can be amplified with the same set of primers used for the clonal tracking library. To ensure that each cell contained a single integrant, NB4 cells were infected at low MOI (5-10% infected cells). To obtain a pure population for each spiked-in barcode, we selected clones with puromycin. The correct number of cells was obtained by serial dilution and added to the leukemic blasts population before extraction of the genomic DNA.

After sequencing and perfect macht alignment to the known spike-in barcodes, for each experimental dataset, we performed a regression analysis to establish the relationship between the number of reads obtained from the sequencing of each spiked-in barcode and the number of cells from which the spiked-in sequences had been obtained (Figure 18 and Table 9).

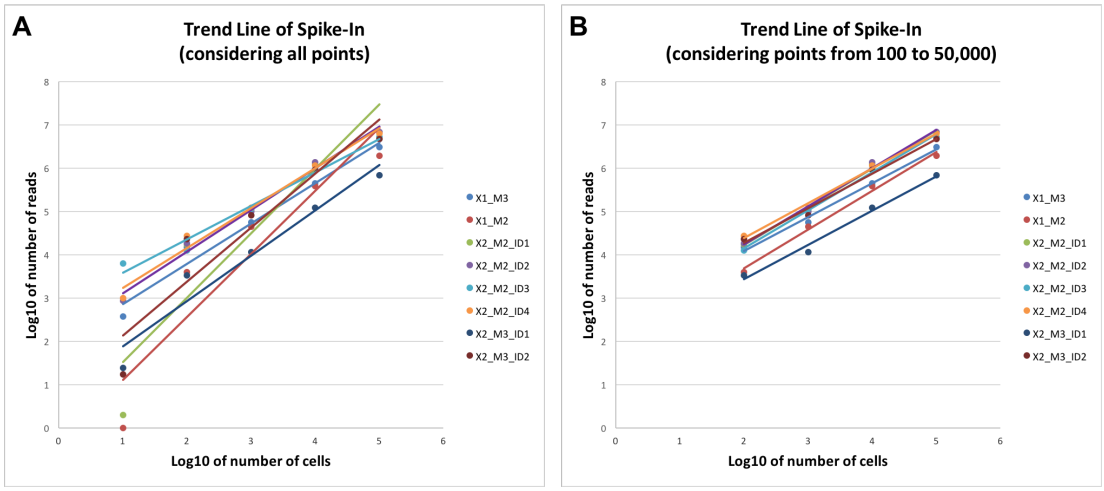

Figure 18. Regression line in spike in experiments. The relationship between the absolute number of cells and the number of reads obtained by HTS for each spike-in control is linear. This relationship was measured for each experimental sample of clonal tracking performed on murine leukemias. The solid lines indicate the linear regression that fits the data for each experiment (each colour identifies a different sample). Values are expressed in logarithmic scale (Log_{10}). **A.** Linear regression obtained considering all dilutions of the spike-in control cells. **B.** Linear regression obtained eliminating the spike-in control equivalent to 10 cells, which, for some of the sample, clearly appears to be out of the linear range.

As shown in Figure 18, for each experiment we observed a linear relationship between the numbers of reads obtained from the sequencing of each spiked-in barcode and the numbers of cells added as spike-in control to the genomic DNA. Even if we were able to identify also the barcodes corresponding to 10 cells in 7 out of 8 samples analysed, for 4 samples this data point was clearly out of the determined linear range (Figure 18A and Table 9). These data suggests, within our experimental conditions, a lower limit of detection of barcodes at around >10 cells. Indeed, considering only the spike-in controls from 100 to 50,000 cells we observed a perfect linear correlation with R^2 grater than 0.99 between number of reads and number of cells (Figure 18B and Table 9), indicating that the number of reads is directly proportional to the number of infected cells harbouring a specific barcode in the tumor population. Moreover, this linear relationship is maintained also for barcodes identified by a very high number of reads (up to ~7 millions), suggesting that our protocol for amplification of the library does not affect the relative frequency of each barcode within the population.

After this analysis, we decided to set the threshold for identification of the clones within the tumor population to a minimum of 10 cells. Therefore, based on the linear regression of the spike-in, for each sample, we calculated the number of reads that correspond to 10 cells and used this number to set the threshold of the background and determine the clonal composition of each murine tumor (Table 9).

Table 9. Number of reads obtained by sequencing (obtained after HTS) or calculated by regression analysis (Calculated) for the spike-in barcodes. For each number of cells for the spike-in controls is reported the number of reads obtained by HTS. Considering that, for at least 4 out of the 8 experiments reported, the spike-in barcode equivalent to 10 cells has a number of reads clearly out of the linear range, we determined the expected number (Calculated) of reads for 1 and 10 cells, considering the linear regression parameters after removal of the 10-cells data points (see panel B of Figure 18). R^2 is the correlation coefficient that measures the reliability of the linear relationship between our two variables. R^2 tot, coefficients obtained considering all data points (panel A of Figure 18). R^2 adj, coefficients of the linear regressions after removal of the 10-cells data points (panel B of Figure 18).

Sample	Number of Reads/Barcodes								
	Calculated		Obtained after HTS						
	1	10	10	100	1000	10000	50000	R^2 tot	R^2 adj
X1_M3	214	1,549	370	14,892	55,648	445,289	3,095,749	0.87082	0.9921
X1_M2	45	437	0	4,017	45,016	387,987	1,925,170	0.97309	0.99169
X2_M2_ID1	170	1,585	2	17,100	113,476	1,233,100	6,633,311	0.85605	0.996
X2_M2_ID2	174	1,622	871	18,488	101,221	1,361,270	6,830,331	0.98919	0.99139
X2_M2_ID3	141	1,318	6,252	12,497	114,384	949,392	5,358,466	0.97942	0.99697
X2_M2_ID4	380	2,884	1,004	27,494	120,795	1,167,220	6,292,147	0.98223	0.99397
X2_M3_ID1	218	1,557	24	3,378	11,426	122,388	681,267	0.94168	0.98671
X2_M3_ID2	151	1,083	17	22,989	81,168	907,633	4,717,180	0.88229	0.98707

As described in the previous paragraph, for the first passage (X1) of mAPL18, we analysed the spleen cells from two independent recipient mice, Mouse 2 and Mouse 3. After HTS and perfect-match alignment, we obtained 1.89×10^7 and 1.42×10^7 total reads, respectively, accounting for 290 different barcodes for Mouse 2 and 226 for Mouse 3 (Table 10). Applying the threshold determined through the spike-in controls, for each sample, we considered as "real" traceable clones only the clones identified by barcodes with a number of reads ≥ than the number of reads calculated by linear regression to correspond to 10 cells: 437 and 1,549 reads for Mouse 2 and Mouse 3, respectively (Table 10). Applying this threshold we obtained 40 and 21 different clones for the two X1 samples. The relative frequency of these clones ranges from 0.002% to 49.20% for Mouse2 and from 0.012% to

42.05% for Mouse3. However, the total number of clones above the threshold covers more than 99.9% of sequenced reads.

Applying the same threshold described for the X1 passage to each X2 sample, again, we always obtained a number of clones (from 11 to 26) that covers more than 99% of the reads (Table 10).

Table 10. Clonal composition of the leukemia mAPL18. For each passage (X) of leukemia transplantation are described: the total number of reads obtained by sequencing, the total number of barcodes identified, the number of barcodes present in more than 10 cells, the percentage of sample analysed and the frequency (%) of the less represented and of the most represented clone. It is worth underlying that, for each leukemia, more than 99% of the sample was analysed. SPL, spleen; ID, mice from which the tumor cells are derived for transplantation as described in Figure 12.

Passage	Cells transplanted	Sample Analysed	Organ Analysed	Total Reads ($x10^6$)	Total Barcodes	Barcode Founded ≥10 cell	% of samples analysed	% lowest clone	% highst clone
X1	Infected cells	ID2	SPL	14.2	226	21	99.91	0.012	42.056
		ID3	SPL	18.9	290	40	99.98	0.002	49.214
X2	Spleen (from X1 ID2)	ID2.1	SPL	33.5	338	26	99.71	0.061	76.371
		ID2.2	SPL	30.9	157	19	99.75	0.062	76.508
		ID2.3	SPL	32.01	164	21	99.86	0.046	78.918
		ID2.4	SPL	28.6	167	11	99.42	0.164	78.16
	Spleen (from X1 ID2)	ID3.1	SPL	4.14	124	13	99.78	0.041	82.124
		ID3.2	SPL	4.72	244	20	99.87	0.039	78.16

For the experiments of clonal tracking in the AML-IEO20, the spike-in internal control was not available, therefore, in order to establish a threshold for clone identification, we decided to take advantage of the data obtained for the spike-in controls of the murine experiments. We observed that, for the murine leukemia, independently from the total number of reads obtained and from the number of reads used to establish our threshold, the frequencies of the less represented clones in the tumors were always very similar in the majoriy of the samples with a mean average of 0.04% (excluding the highest (0.164%) and the lowest (0.002%) values) (Table 10, % lowest clones). Based on our definition of threshold for the murine leukemia, we can extrapolate that the frequency of 0.04% corresponds to the frequency of the clones represented by 10 cells in the tumor. We used these parameters to establish a threshold also in the human samples. Therefore, in order to

study clonal evolution *in vivo*, for each passage of hAML-IEO20, we considered and compared exclusively the barcodes with a frequency ≥0.04%.

As for the murine leukemia, we observed that, in each sample analysed, the majority of the clones were present with a very low frequency and that only few tens were above the threshold of 0.04%. However, for all samples analysed for AML-IEO20, it is clear that all clones with a clonal frequency ≥0.04% cover more than 99% of sequenced reads. Table 11 shows the results for the clonal analysis of serial transplantation of the leukemia derived from Mouse 2 and Table 12 for the leukemia derived from Mouse 4.

Table 11. Clonal composition of the AML-IEO20, Mouse ID2. For each step (X) and each organ analyzed are described: the total number of reads obtained by HTS, the total number of the barcodes identified, the number of barcodes with a frequency ≥0.04%, the percentage of tumor sample analysed and finally the percentage of the most represented clone. Considering only clones with a frequency ≥0.04%, more than 99% of the tumor is always analyzed. BM, bone marrow; SPL, spleen; ID, mice from which the tumor cells are derived for transplantation as described in Figure 12.

Passage	Transplanted cells	Sample Analysed	Organ Analysed	Total Reads ($x10^6$)	Total Barcodes	Barcodes with freq≥0,04%	% of samples analysed	% highest clone
X1	Infected cells	ID2	BM	107	3413	66	99.30	18.75
			SPL	94.6	5236	87	99.02	16.54
X2	Spleen (from X1 ID2)	ID2.1	BM	66.4	329	27	99.88	34.21
			SPL	28.5	317	32	99.85	29.8
		ID2.2	BM	44.8	318	16	99.76	61.44
			SPL	41.8	424	22	99.87	71.85
	Bone marrow (from X1 ID2)	ID2.3	BM	85.1	430	15	99.76	43.37
			SPL	98.0	532	25	99.87	46.56
		ID2.4	BM	54.2	227	23	99.82	43.67
			SPL	35.2	199	21	99.84	40.34
X3	Spleen (from X2 ID1)	ID3.1	SPL	1.76	84	8	99.94	72.23
	Bone marrow (from X2 ID1)	ID3.2	SPL	1.30	81	4	99.81	52.32
	Spleen (from X2 ID4)	D3.3	SPL	7.12	69	9	99.86	44.72
	Bone marrow (from X2 ID4)	ID3.4	SPL	2.97	78	4	99.60	41.53

Table 12. Clonal composition of the AML - IEO20, Mouse ID4. As for Mouse 2, for each animal transplanted with AML-IEO20 are described: the total number of reads obtained by HTS, the total number of barcodes identified, the number of barcodes with a frequency ≥0.04, the percentage of tumor sample analysed and, finally, the percentage of the most represented clone. As for Mouse 2, applying a threshold of clonal frequency ≥0,04% more than 99 % of the tumor is analyzed in most of the samples. BM, bone marrow; SPL, spleen; ID, mice from which the tumor cells are derived for transplantation as described in Figure 12.

Passage	Transplanted Cells	Sample Analysed	Organ Analysed	Total Reads ($x10^6$)	Total Barcodes	Barcodes with freq≥0,04%	% of samples analysed	% highest clone
X1	Infected cells	ID4	BM	41.7	1539	96	98.98	16.82
X2	Bone marrow (from X1 ID4)	ID2.1	SPL	49.2	157	17	99.89	29.17
		ID2.2	SPL	12.5	171	13	99.94	34.26
		ID 2.3	SPL	20.2	60	5	96.72	31.42
		ID2.4	SPL	33.8	113	11	99.9	24.68
		ID2.5	BM	83.9	595	22	99.47	31.14
X3	Bone marrow (from X2 ID4)	ID3.1	BM	1.96	65	8	99.62	47.99
		ID3.2	BM	3.03	54	8	99.9	49.33

4.3.3 The clones defined by the clonal tracking experiments identify leukemia initiating cells (LICs)

Up to date, the gold standard to measure the frequency of LICs, and of HSCs in general, is the limiting dilution transplantation (LDT) assay. In this assay, a serial dilution of cell doses of the leukemia are transplanted in several recipient mice. The frequency of LICs is then estimated based on the ratio between the total number of mice transplanted for each cell dose and the number of mice positive for reconstitution at each cell dose, using the web tool Extreme Limiting Dilution Analysis (ELDA)[98]. This software calculates a range of stem cells frequency within the tested cellular population, with a confidence interval of 95%. Thus, we calculated LIC frequencies of the leukemias used for clonal tracking experiments, using the LDT assay. As listed in Table 13, the stem cell frequency determined by ELDA is 1/989 (confidence interval: lower 1/1844; upper 1/530) and 1/471 (confidence interval: lower 1/838; upper 1/265) for AML-IEO20 and mAPL18, respectively.

Table 13. Estimation of LIC frequency by LDT assays. Experiments of limiting dilution transplantation were performed both for AML-IEO20 and mAPL18. The frequency of stem cells was calculated using the software ELDA, based on the number of animals that developed leukemias out of the total number of transplanted animals for each cell dose.

human AML-IEO20		
# of cells per mouse	**# of mice transplanted**	**# of dead mice**
500,000	3	3
100,000	5	5
50,000	4	4
10,000	8	8
5,000	9	9
1,000	11	6
500	5	3

Confidence intervals for 1/(stem cells frequency)		
Lower	**Estimate**	**Upper**
1,844	**989**	530

murine APL #18		
# of cells per mouse	**# of mice transplanted**	**# of dead mice**
1,000,000	5	5
200,000	9	9
100,000	10	10
50,000	9	9
20,000	7	7
10,000	15	15
5,000	0	10
500	18	11
50	8	2

Confidence intervals for 1/(stem cells frequency)		
Lower	**Estimate**	**Upper**
838	**471**	265

The experimental conditions of the LDT assay are quite different from the experimental conditions used in our studies, in which we transplant the leukemic population in bulk. We decided to compare the two methods for LIC determination in order to assess if the barcoding strategies and the threshold we defined for our study are reliable in terms of identification of LICs. We, therefore, determined the LIC frequency for AML-IEO20 and mAPL18 based on the results of clonal tracking experiments. In particular, we estimated the LICs taking into consideration the percentage of barcoded cells in the transplanted leukemic population (% of infection) and the average number of clones identified in the passage X1 for each leukemia (barcodes above the set thresholds), applying the following formula:

$$LIC\ frequency = 1{:}\left[\#of\ Transplanted\ cells{:}\left(\frac{Barcodes\ > 10\ cells * 100}{\%\ of\ infection}\right)\right]$$

For human AML-IEO20, for all three primary samples analysed, the LIC frequency calculated based on the results of the clonal tracking experiments are within the confidence

intervals calculated by ELDA and the average LIC frequency is 1:938, virtually identical to the one of 1:989 estimated by ELDA (Table 13 and 14). For the murine APL, the LIC frequencies calculated based on our clonal experiments were slightly below the range determined by LDT (1:838), with an average frequency of 1:1452 (Table 14).

Our data show that, in spite of the different behaviours growth patterns of the different leukemic samples, the LDT and the clonal tracking experiments give similar results, indicating that the clones identified through our experimental strategy identify *bona fide* LICs.

Table 14. Comparison of the LIC frequencies. The LIC frequency were calculated by ELDA or based on the number of clones identified in the passage X1 of our clonal tracking experiments.

Leukemia	Sample Analysed	Organ Analysed	# of clones in X1	LIC fesquency by ELDA	LIC fesquency by Clonal Tracking
Human AML-IEO 20	ID2	Spleen	87	1:989	1:862
	ID2	Bone marrow	64		1:1,172
	ID4	Bone marrow	96		1:781
Murine APL #18	ID2	Spleen	40	1:471	1:1,000
	ID3	Spleen	21		1:1,904

4.3.4 The clonal composition of the leukemic blasts in the different hematopoietic compartments is highly similar

We first asked if there is a difference in the clonal composition of leukemic blasts collected from different hematopoietic organs, e.g. if there is selection of different clones within different hematopoietic compartments. To this end, we performed clonal tracking analysis of blasts derived from spleen and bone marrow of the human AML-IEO20 from the leukemia transplanted in Mouse 2. In particular, we analysed the blasts derived from spleen or BM of the primary recipient (X1) and from 4 independent secondary recipients (X2) (Figure 19).

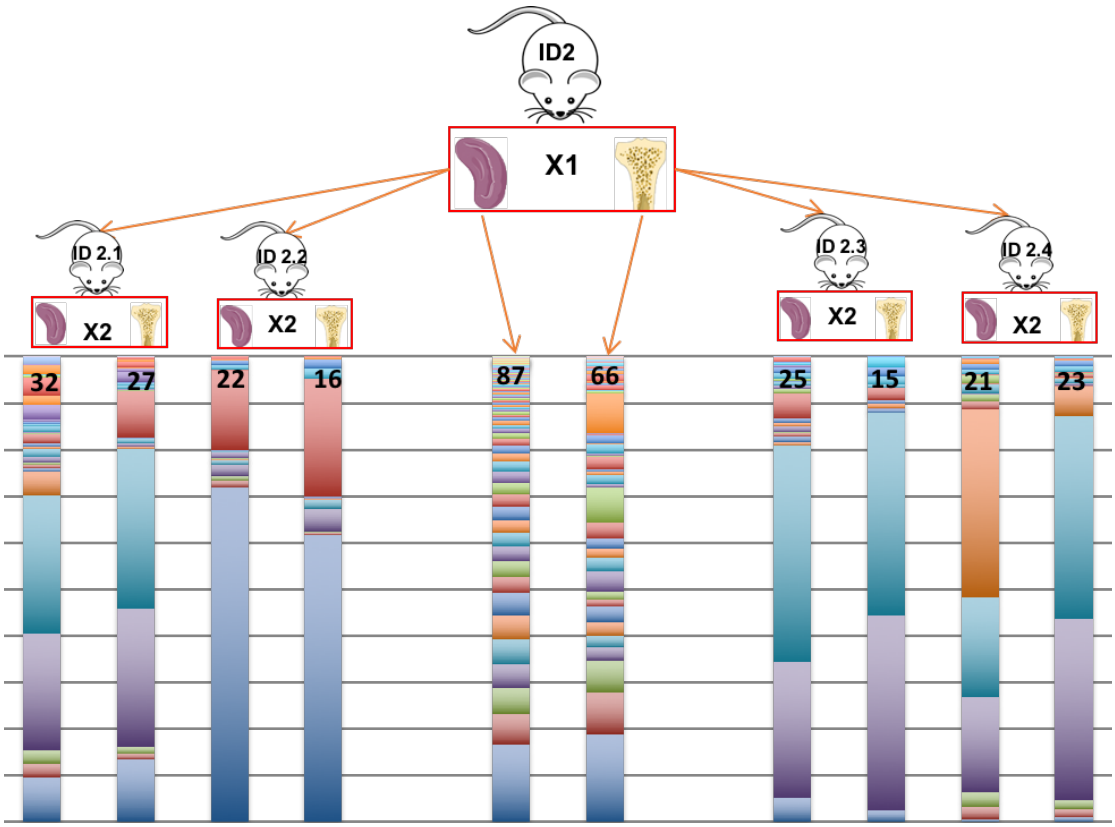

Figure 19. Clonal evolution of blasts derived from spleen and BM of human AML-IEO20 (Mouse ID2) in two passages of serial transplantation. Leukemic blasts collected from spleen (left bars for each mouse) and bone marrow (right bars for each mouse) of the X1 mouse (ID2), were transplanted in 4 secondary recipients (X2). At leukemia development in the secondary recipients, barcodes analysis was performed from spleen and bone marrow of the 4 independent animals. The graphic shows the number of clones identified in the two organs of each mouse. Each bar of different colour identifies a different clone (barcode). The height of each bar indicates the frequency of each clone in the tumor population. The clones are ordered bottom to top with decreasing frequency. Each colour corresponds to a specific clone and the colour code is maintained for all bars.

As shown in Table 11 and Figure 19, the number of clones identified was very similar not only between spleen and BM extracted from the same animal but also between the two compartments from two independent recipients.

Notably, the clonal composition within the two compartments was very similar both in terms of clonal identity and clonal frequency among all recipients (Figure 20). Indeed, the correlation of the clonal distribution between spleen and BM was very high (R=0.86) for the X1 passage of transplantation and increased further in the second passage of transplantation, with a Pearson correlation coefficient >0.95 for all X2 samples, with the exception of one mouse (ID2.2), which showed a different clonal composition, as compared to all other animals (Figure 26).

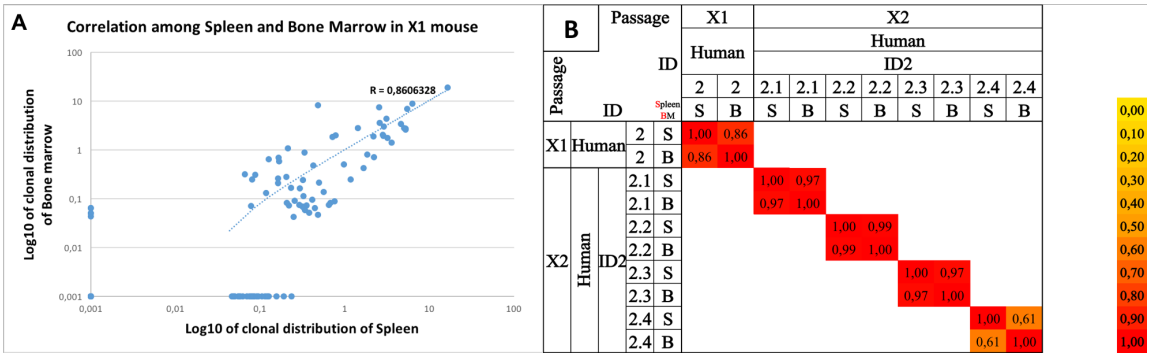

Figure 20. The clonal composition of the leukemic blasts in spleen and BM is highly similar. A) Clonal distribution of the leukemic clones in the spleen (*x* axis) and in the BM (*y* axis) in the X1 mouse ID2 transplanted with AML-IEO20. The scatterplot shows the frequency distribution of clones expressed in logarithmic scale (Log_{10}). Each solid circle identifies a clone. In the middle of the plot are clones in common between the spleen and BM. On the x axis the unique clones for the spleen and on the y axis the clones unique for the BM. B) Heatmap of the R scores determined by Pearson correlation analysis. For each pair of samples we calculated the correlation of the distribution of the identity and frequency of the clones identified in the spleen (S) and BM (B) of each indicated sample according to the experimental scheme and data depicted in Figure 12. ID number, mice from which the tumor cells are derived for transplantation; X, passage of transplantation.

Our data indicate that there is no selection of leukemic clones within the different leukemic organs and that there is no clonal tropism for the distribution of the disease. Based on the homogeneity of the clonal analysis observed from the two hematopoietic organs, also in order to reduce the costs of our analysis, we decided to consider (in the subsequent clonal tracking analyses) only one hematopoietic compartment *per* recipient mouse (BM or spleen indifferently).

In the following clonal tracking analysis, we considered three parameters: i) the number and the identity of the clones among the different transplanted mice and the different passages of leukemia transplantation, searching for the clones in common; ii) the relative distribution of the frequency of the clones within the tumor population, independently from their identity: we call this parameter *"correlation of frequency"*; iii) the frequency of each individual clone within the tumor population: *"correlation of identity"*. We performed the analysis of these three parameters both "horizontally", within the different animals transplanted with the leukemias within the same passage of transplantation, and

"vertically", looking at how these parameters varied with the serial passages of transplantation.

4.3.5 The number of clones able to engraft and their ability to expand is very similar at each passage for all leukemia analysed

We first try to understand the behavior of the same AML sample after transplantation in different animals and at different passages, by analysing the numbers of clones that are able to engraft for each leukemia at each passage. As shown in Figure 21 and reported in Tables 10, 11 and 12, the number of clones identified for each sample is very similar at each passage. This indicates that there is a limited and fixed number of LIC able to engraft and propagate the disease.

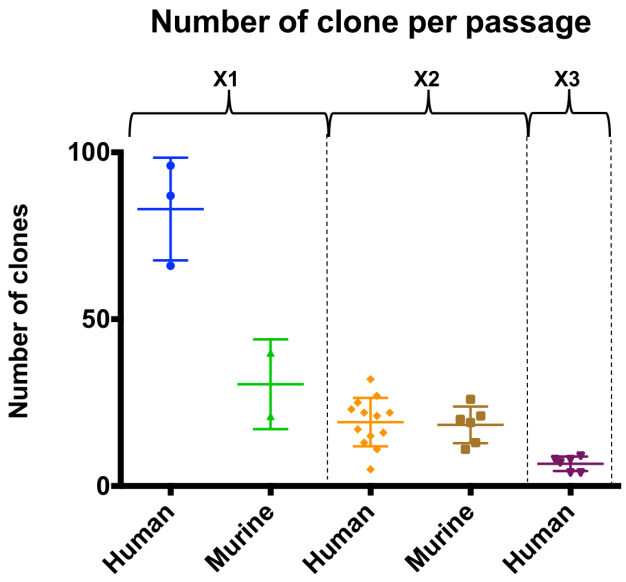

Figure 21. Number of clone per passage. For each passage the total number of engrafted clones identified above threshold were considered.

Moreover, we noticed that within each passage, independently from their molecular identity, the distribution of the frequency of the clones within each transplanted animal was highly conserved among all the different animals, for both AML-IEO20 and mAPL18 (Figure 22).

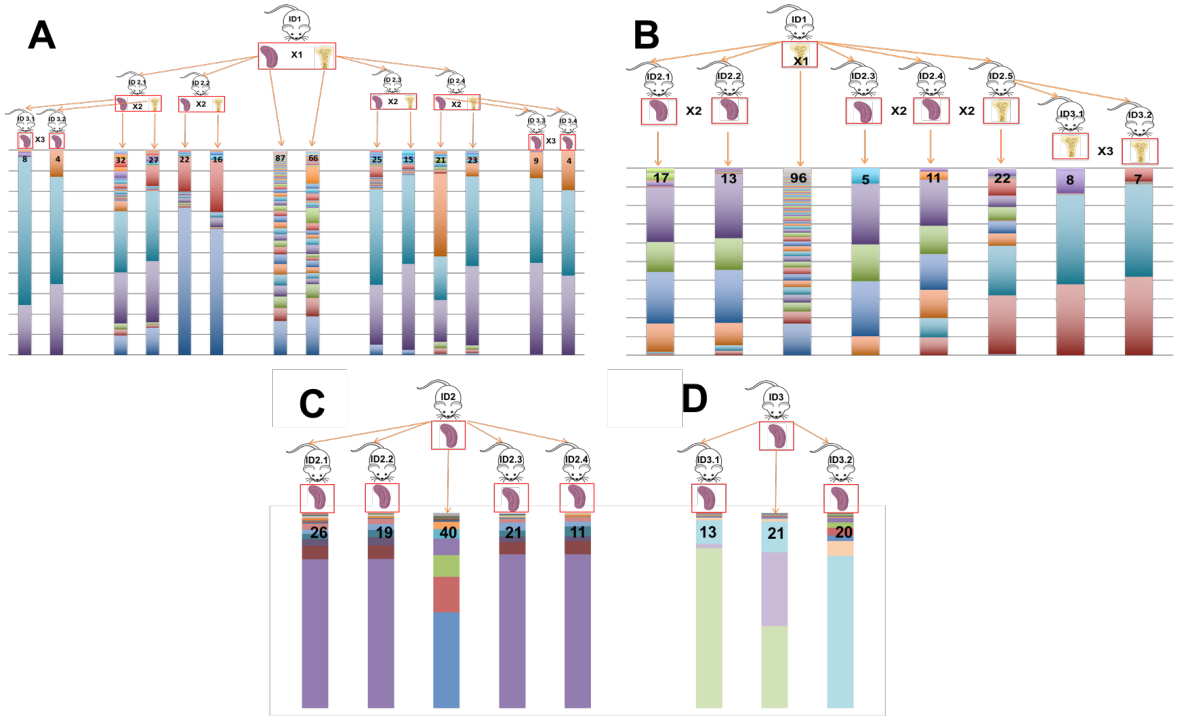

Figure 22. Clonal tracking experiment: A) human AML-IEO20 ID2, B) human AML-IEO20 ID4, C) murine APL18 ID2 and D) murine APL18 ID3. As described in Figure 19 each colour corresponds to a specific clone and the colour code is respected for all serial passages.

We then used the Pearson correlation coefficient (R) to measure the degree of conservation of "clone *frequency"* among all transplanted mice within each passage. The *correlation of frequency* was very high even among mice transplanted with different leukemias and in all three passages of serial transplantation, as shown in the heatmap of the Pearson correlation (Figure 23). All Pearson correlation coefficients are above 0.66 with a mean of 0.88. This degree of conservation was particularly high (R >0.9): i) for all primary samples, independently of their human or murine origin; ii) for most of the human secondary and tertiary recipients, even if coming from different primary donors (human ID2 vs ID4); iii) for all secondary recipients of the murine leukemia, for which there was perfect identity of the *correlation of frequency* among all samples (R=1.0).

Taken all together, our data suggest that the number of LICs and their ability to expand within the tumor is somehow determined and finite.

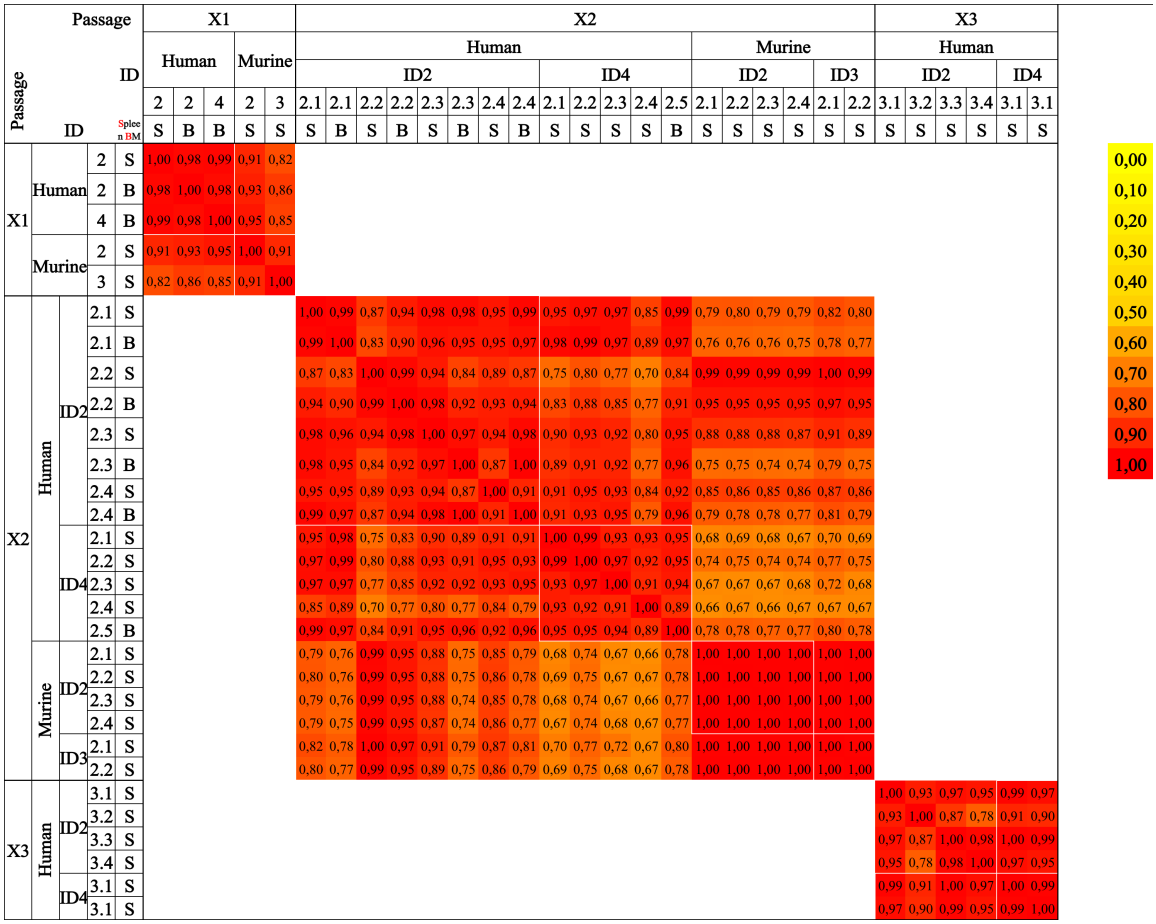

Figure 23. Heat map of correlation of frequency distribution. The correlations were ccalculated taking in consideration the distribution of frequency in each sample compared with the other ones from same passage.

4.3.6 The clones that mainly contribute to leukemia development are identical and have similar growth potential.

Considering the high degree of conservation, of the clones within all leukemias, both in terms of number and frequency distribution, we investigated if the clones that grow with the same frequency within each tumor have also the same molecular identity. In other words, we searched for clones in common among the different mice transplanted with the same leukemia. For each passage of transplantation, we determined the percentage (%) of clones in common among mice transplanted with cells from the same passage of transplantation. We could apply this analysis only to secondary and tertiary passages of transplantation, because the X1 samples derive from independent infections of pools of blasts that, according to our experimental condition, do not share any barcode (Figure 24).

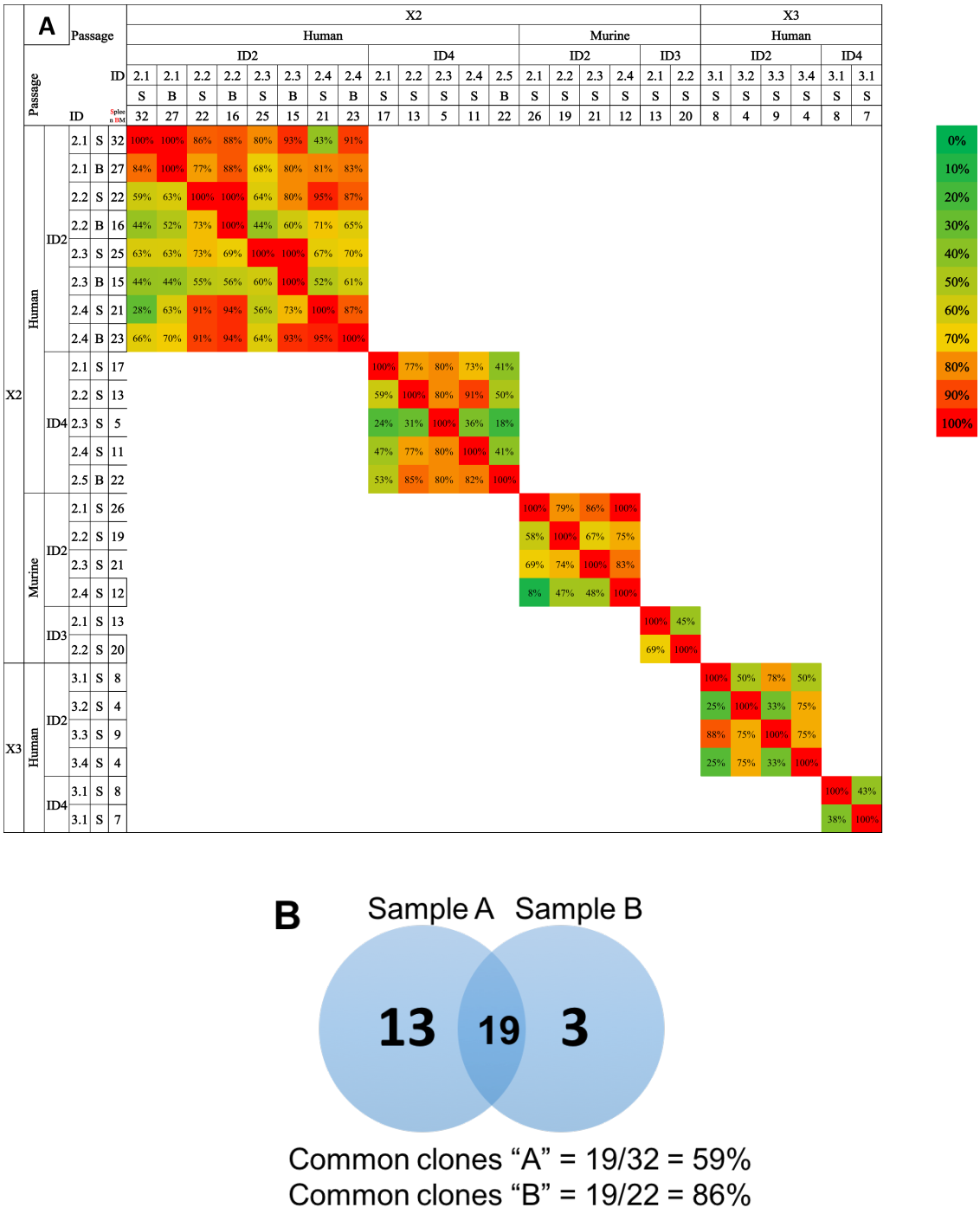

Figure 24. Heatmap of percentage of common clones. A) Heatmap of percentage of common clone among mice from same passages. B) Each percentage of common clones is calculated considering both samples.

As shown in Figure 24, we observed that the percentage of common clones among the different samples was variable, ranging from the 8 to 95%, with ~40% (23/58) of all the

analysed combinations showing <55% of common clones. Interestingly, this was true also for those samples showing a very high *correlation of frequency* (Figure 23). As an example, the murine leukemia showed a Pearson correlation coefficient equal to 1 for every comparison, and less than 50% of common clones, with two samples sharing only 8% of common clones. Likewise for the X3 passage of the human leukemias, which showed *correlation of frequencies* of ≥0.8 for each combination, and <50% of common clones in most samples (8/14; ~60%) (Figure 24). These data indicates that, for most of the samples, the clones that contribute to leukemia development vary within the same leukemia and among different passages. Indeed, the numbers of unique clones for all leukemic samples analysed at each passage of transplantation tend to be higher for both X2 and X3 passages (Figure 25).

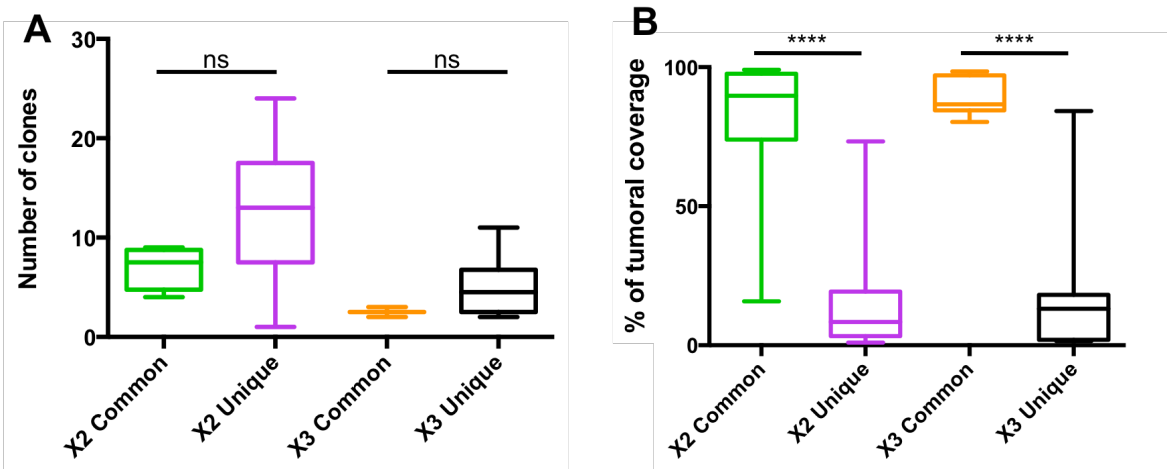

Figure 25. Numbers and relative frequencies of common or unique clones. A) We plotted the numbers of clones (common or unique) found among several mice from the same passage transplanted with the same cells. Common clones tend to be lower compared with the unique. B) in this panel were considered the percentage of tumor that clones decribed in panel A cover.

However, if we consider the percentage of tumor to which the common and the unique clones contribute to, we observe a statistically significant inverse correlation, suggesting that at each passage of transplantation the common clones contribute the vast majority of the tumor (mean of 80.9% and 89% for the X2 and X3 passage, respectively; Figure 25B).

As a consequence, analyses of the distribution of frequency within the tumoral population exclusively for the common clones (socalled *correlation of identity*) revealed that the vast majority of the samples of both human and murine leukemias are identical, with a Pearson correlation coefficient >0.9 (Fig 26). The only exceptions are one sample for each human leukemia: ID 2.2 for Mouse 2 (R ~ -0,138) and ID2.5 for Mouse 4 (R ~ -0,75), that clearly show the expansion of different clones compared to the other samples (Figure 19 and 22).

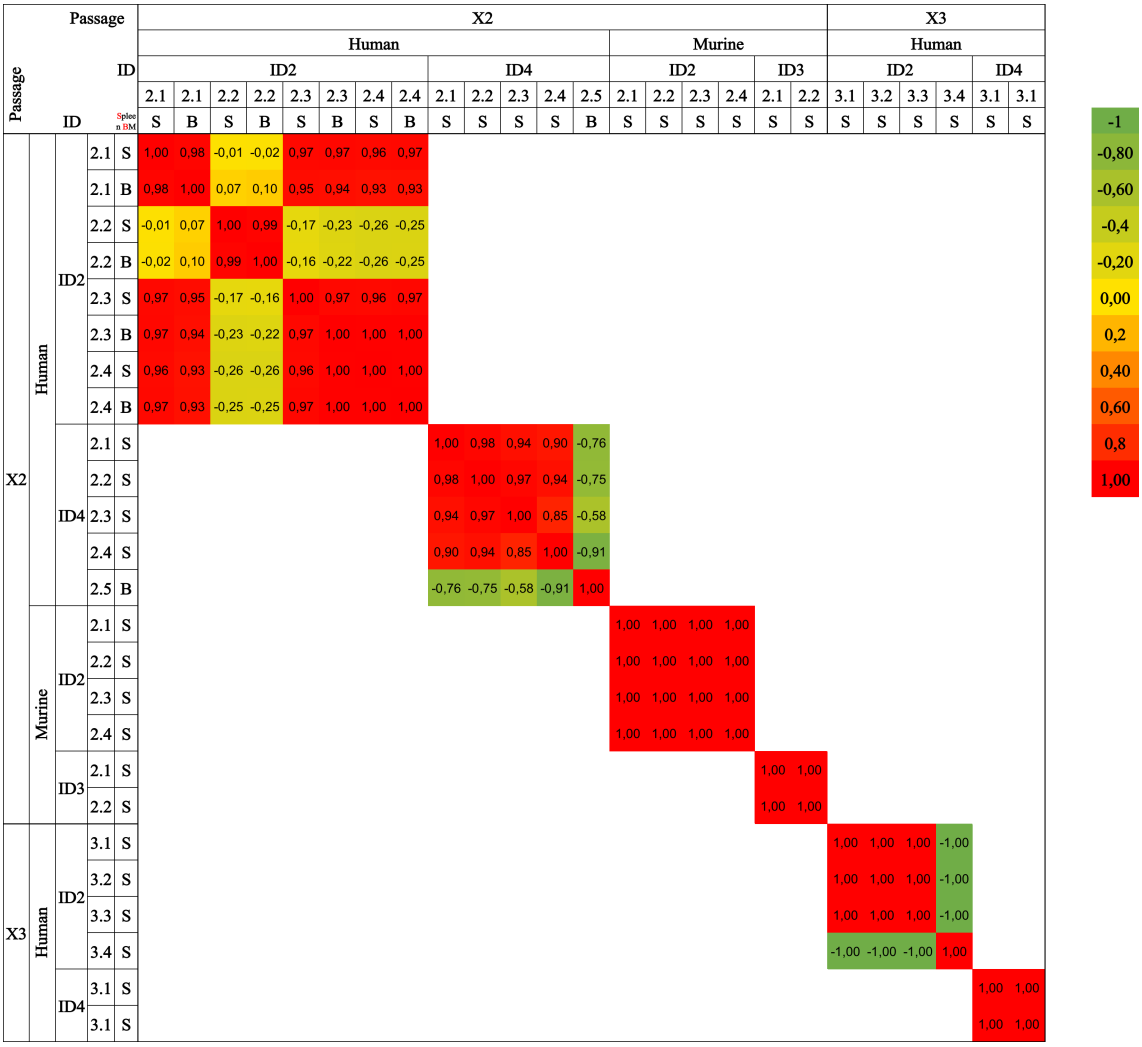

Figure 26. Heat map of correlation of identity. Pearson correlation were calculated among common clones founded in each passage.

Our data indicate that, with the exception of rare cases, for the leukemic samples analysed, the vast majority of each tumor is composed of the same identical clones and that each clone maintains its growth potential upon transplantation.

4.3.7 A clonal selection exists through evolution of the leukemia in the different passages

We next investigated the pattern of evolution of AMLs, comparing the clonal composition of each sample from the X1 to the X2 and to the X3 passage of transplantation within populations of samples derived from the same primary donor. We first analysed numbers of clones at each passage of transplantation. Figure 27 shows that the numbers of clones decreases at each passage of transplantation with a mean of 62±31 for X1, of 19±7 for X2 and of 7±2 for X3 leukemias. The decrease in clonal number is statistically significant among all passages, clearly indicating that each passage is associated with a strong clonal selection.

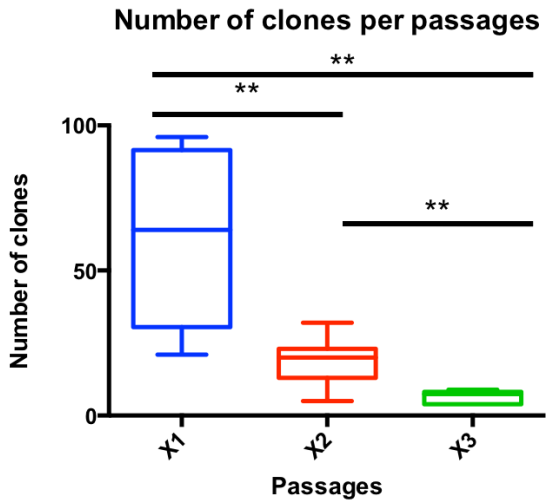

Figure 27. The number of clones decreases through the passages. The boxplots show the total number of clones identified for each passage of transplantation (Passages, X). Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.

In spite of the clonal selection within sequential passages, if we determined the *correlation of frequency* among all tumors through the passages, once again, we scored a very high correlation rate. R scores were never <0.6 and the mean R score was 0.88 (Figure 28),

suggesting the clones that survive serial transplantation have the same potential to populate the tumor.

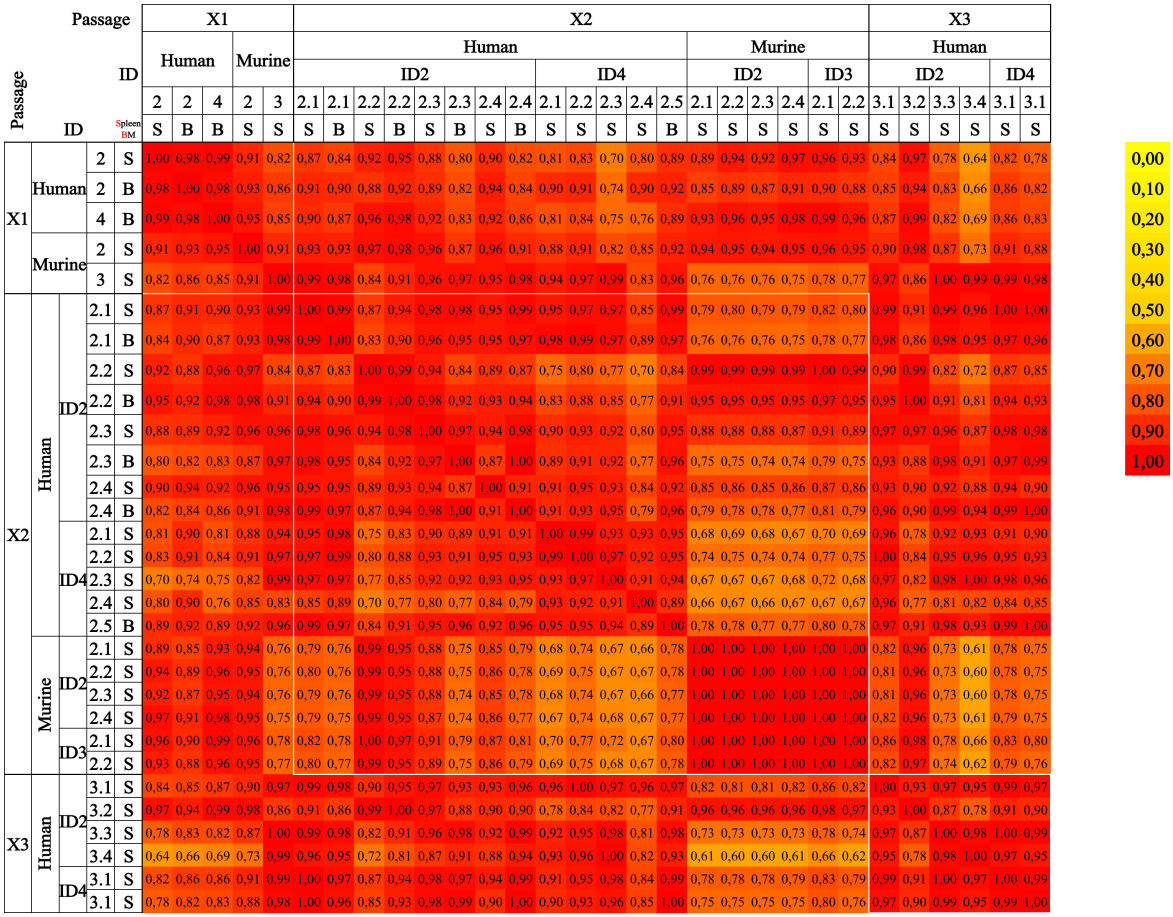

Figure 28. The correlation of frequency among all tumors through the passages is high. Heatmap of the R scores determined by Pearson correlation analysis. For each pair of samples we calculated the correlation of the distribution of the frequency of the clones identified.

We next investihgated the molecular identity of the clones undergoing selection. First, we noticed that all clones of the X2 passages are also present in X1, and all clones of X3 are present in X2, with no new clones appearing in the subsequent passages. The only exception to this situation is mAPL18. Indeed, for this leukemia, in X2 we identified new clones not present in X1. However, these new clones are few, only 10 and 7 for Mouse ID2 and ID3, respectively, and, more importantly, they contribute to a very low percentage of the tumor (≤ 0.35%, Figure 29), suggesting that they might have escaped detection in X1.

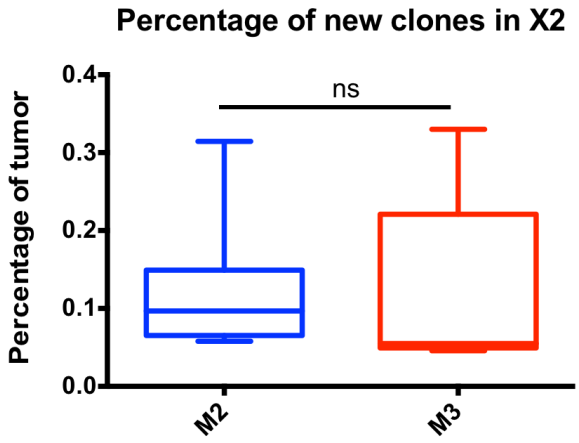

Figure 29. The new clones appearing in mAPL18 are present at very low frequency in the X2 tumor. The boxplots show the distributions of the percentage of tumor contributed by the new clones identified in the passage X2 of Mouse 2 (M2) and Mouse 3 (M3) of mAPL18. The contribution of these clones in X2 is always below 0.4%. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.

We then investigated if there is a correlation between frequency of each given clone within the tumor population and its disappearance at the subsequent passage. To this end, we compared the distribution of the frequencies of clones specifically present only in X1 (X1-specific), with that of clones common between X1 and X2. Same analysis was performed for the X2-X3 passages. For the X1-X2 passages, results showed no differences among the two groups of clones, indicating that low and high frequency clones have equal probability to survive the X1-to-X2 passage (Figure 30A). In contrast to what we observed for the X1 passage, in the analyses of the X2-X3 passage, the clones specific only for X2 have a significant lower frequency in the tumor population compare to the one in common with X3 (Figure 30B). This indicates that in the passage from X2 to X3 passage only the most represented clones were able to survive transplantation.

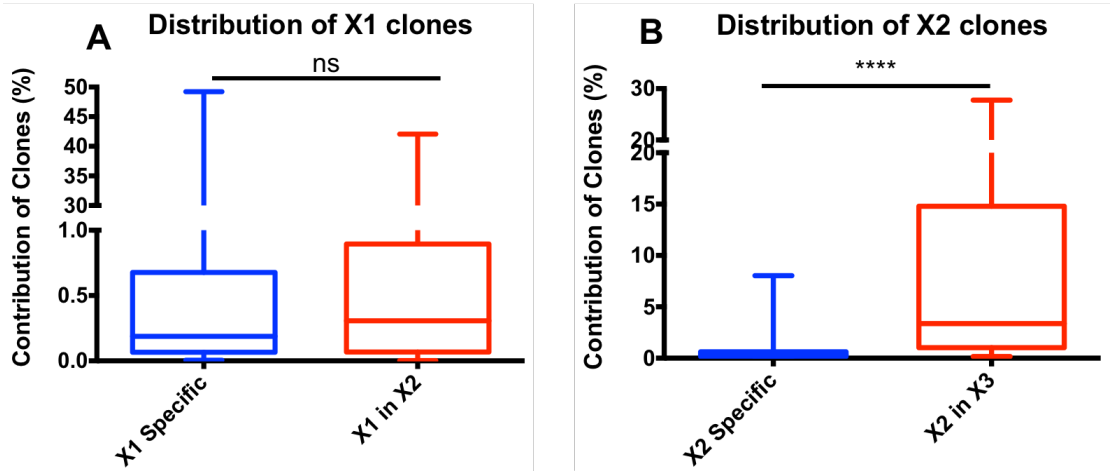

Figure 30. The counter-selected clones are not necessarily the clones present at the lowest frequency within the tumor population. A) Distribution of the frequencies within the tumor population of the X1 clones present exclusively in the X1 passage of the leukemia (X1 specific) versus the distribution of clones in common with X2 passage (X1 in X2). B) As for panel A for the clones present in the X2 passage of the leukemia: X2 clones present exclusively in X2 samples (X2-specific) and in common with X3 (X2 in X3).

Indeed, these data were confirmed also by the analysis of the percentage of tumor that the common clones contribute to in the previous passage of leukemia transplantation: while the clones in common with X2, in X1, covered a wide range of percentages of the tumor population, the clones in common with X3, in X2, contributed only to a very high percentage of the tumor (Figure 31A and 31B).

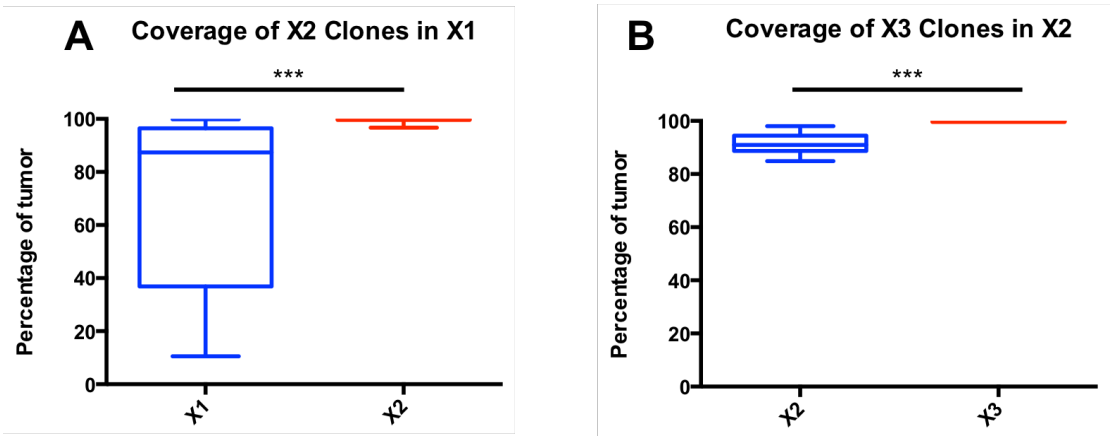

Figure 31. The contribution to the tumor population of the common clones between passages is different between the X2 and the X3 clones. A) Tumor coverage (percentage of tumor), measured as percentage of each clone within the tumor population, for the clones in common between X1 and X2 both in the X1 and in X2 passage of transplantation. B) As for panel A for clones in common between X2 and X3.

Taken all together our data suggest that the clonal selection between passage X1 and X2 appear to be deterministic, therefore not linked to the frequency of each clone within the tumoral population but to intrinsic biological properties of the transplanted cells. In contrast, the clonal selection between passage X2 and X3 appear to be more stochastic, very likely due to a bias in the sampling of the most represented clones that in X2 account for the majority of the leukemic population.

4.3.8 The mode of evolution through the passages of each LIC is an intrinsic cellular property.

We finally assessed how the clones evolve in time through the passages. We first determined which clones specifically appear or disappear completely from the X1 to the X2 passage. In agreement with the clonal selection reported in the previous paragraph, the clones that disappeared completely were in total 137 and represented ~52% of all (265) X1 clones identified (Figure 32A). As described above, the clones that, instead, appeared in X2 are very few (only 17), are peculiar of mAPL18 and reach frequencies within the tumor population below 0.35% (Figure 32B and Figure 29).

We then considered the clones in common between the passage X1 and X2: 110 in total. For these clones we determined the *correlation of identity*, measuring how the frequency of each individual clone within the tumor population varies between passages. According to their behaviour during clonal evolution, we identified three classes of clones: decreasing clones, growing clones and stable clones. The decreasing clones significantly decrease their frequencies in X2, they are 29 and account for 27% of the common clones (Figure 32A). The vast majority of the clones are growing clones (64 clones, 58%), the one that increase their frequency from X1 to X2 (Figure 32B). The stable clones, instead, tend to maintain the same frequency between the two passages. These clones are the less represented (17 clones, 15%; Figure 32C).

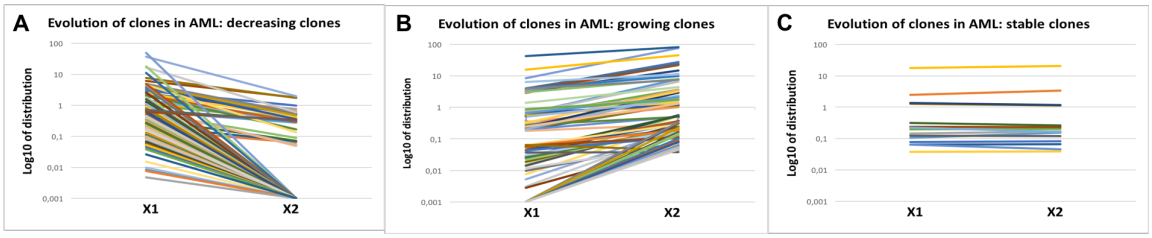

Figure 32. Clonal evolution of AML between X1 and X2. A) Each line identifies a different leukemic clone. The graphs show the frequency distribution of each clones at each passage (X). Based on their frequency within the tumor population in the two passages the clones are divided in 3 classes: decreasing (A), growing (B) and stable (C).

If we, then, looked at the further step of evolution in the X3 passage, the vast majority (36/39, 92%) of the decreasing clones, again, disappeared. Only 3 clones survived in X3 but with lower frequencies compared to X2 (Figure 33A and D), again suggesting a strong clonal selection. Within the growing clones, only 8 constantly increased, expanding from X1 to X2 and expanding again in X3, and reaching frequencies that account for more than 50% of the entire tumor population (Figure 33B and E). We did not score any stable clone within the leukemic population from X2 to X3, however, we noticed the presence of clones with a peculiar behavior. These clones account for half of the common clones (50 %). They increase their frequencies from X1 to X2 and tend to disappear in X3 (Figure 33C). Notably, this behavior was true even for those clones with a frequency up to 30% within the tumor population in the X2 passage (Figure 33F).

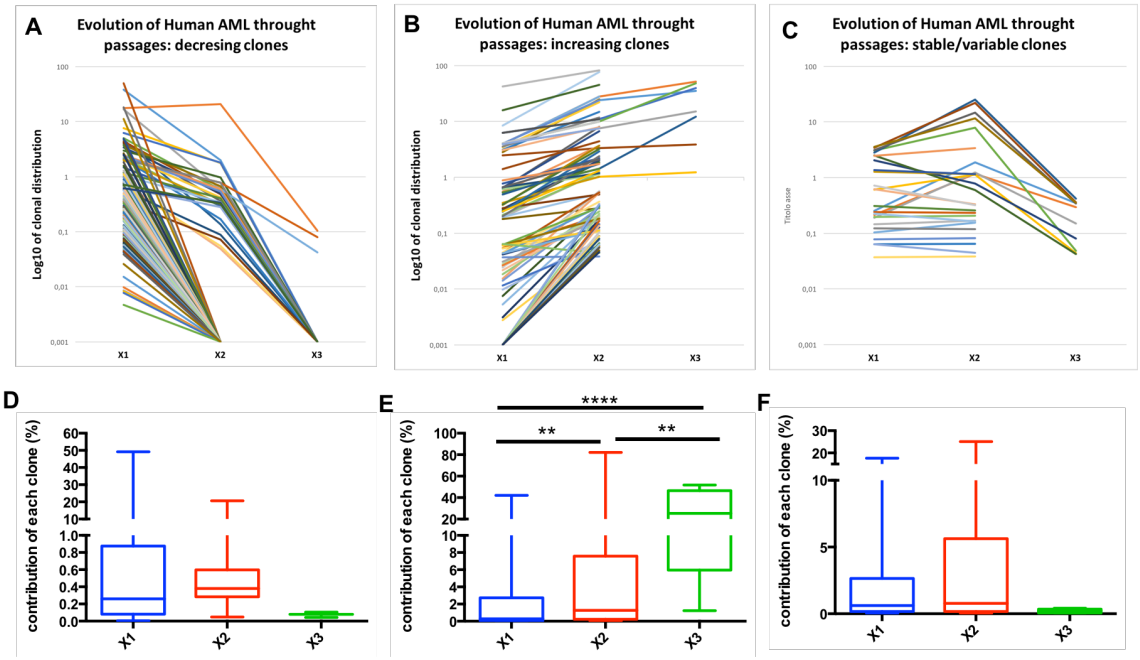

Figure 33. Clonal evolution of AML within the 3 passages of transplantation. A-C) As described in Figure 28. D-E) The boxplots show the frequency distributions within the tumor for each passage (X) of each class of clones: D, decreasing; E, growig; F, stable or variable. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.

Taken all together, our data suggest that the different LICs within the leukemic population have different proliferation properties and can be mainly classified in 3 distinct classes based on their destiny through the passages: LICs that immediately exhaust, LICs that have a limited proliferative capacity and LICs that maintain their indefinite proliferative capacity.

4.4 shRNA experiments

4.4.1 Determination of the maximum complexity of the shRNA libraries supported by our leukemic models

Pooled short hairpin RNA (shRNA) libraries have been successfully used to perform *in vivo* genetic-screens to identify novel oncogenes and tumor suppressors in murine models of liver cancer [25] and lymphoma [26, 27] .

In order to define the leukemogenic potential of several genes in a single assay, we have performed an *in vivo* shRNA screening in xenotransplants of patient-derived leukaemia. Perform genetics screens in primary human tumours *in vivo* is quite innovative and it has not been reported yet. Instrumental for this study was the collection of human leukemic blasts able to grow in immuno-deficient mice and characterized by different and defined genetic alterations. Moreover, since the ability of AML to grow in a mouse model rely on LICs, this approach should, in principle, lead to the identification of crucial pathways for the growth of LICs *in vivo.*

Particularly critical for the *in vivo* shRNA screening is:

1. Determination of the multiplicity of infection, so that each cell is infected with a single viral particle and, therefore, has the integration of a single shRNA in its genome.

2. Determination of the shRNA library complexity that the cancer model can support, or, in other words, how many genes can be efficiently targeted by a pool of shRNA in a given tumor sample.

Regarding the first point, it has been previously published[96] that a percentage of infection below 20% should ensure that each cell is infected with a single viral particle. To determine if this was true also in our experimental system, we conducted a virus titration assay by qPCR on the AML-IEO20 sample infected with the epigenetic library hEPI1 at MOI of 2 that, in our experimental setting, guaranteed that about 10% of blasts got infected. Half of the infected blasts were selected in puromycin (1 μg/mL) for 96 hours,

then, the DNA was extracted and used for the QPCR-based analysis as described in details in the section 3.9.2 of Materials and Methods[96]. Briefly, starting from the qPCR Ct values obtained for the infected sample and the standard curves both for viral and human sequences, we have calculated the number of integrant *per* cell. The results are reported in Table 15.

Table 15. Calculation of number of integrations per cell. The results were obtained by qPCR analysis as reported in Material and Method section 3.9.2. Ct, Cycle threshold. WPRE, Woodchuck Hepatitis Virus (WHP) Posttranscriptional Regulatory Element (WPRE). hAlb, Albumin.

Sample	Ct	Average WPRE	Average hAlb	Integration/cell
AML-IEO20 Not Infected	31.72	77	13,500	0.011
AML-IEO20 Not Selected	24.68	10,900	20,000	1.09
AML-IEO20 After Selection	24.4	13,000	20,500	1.271

For the second point we had to consider that only LICs are able to engraft and proliferate significantly after transplantation into recipient mice[83], therefore, in our experimental design, we should cover the complexity of the library within the stem cell population of the leukemia.

In order to estimate the LIC frequency in our tumor samples, we have performed clonal tracking experiments and/or limiting dilution transplantation assays. In Table 14 is reported the LIC frequency estimated both by limiting dilution and clonal tracking experiments for one human and one mouse AML sample. As evident, these two methods gave similar results, therefore, for the other human samples we have performed only the limiting dilution approach that, for all samples analysed, gave a LIC frequency close to 1 out of 1000 blasts (data not shown). Previous studies have shown that in order to ensure a good representation of all shRNA during the screening *in vivo*, we need to infect at least 5-10 LICs *per* shRNA[86,99]. Considering that the infection of 50 million blasts is our limit of

availability and/or manageability for both leukemic blasts and viral particles, and considering an average infection efficiency between 10 to 20%, we can screen about 1000 different shRNAs. Finally, since in order to avoid off-target effects each gene has to be targeted by 10 shRNAs, our library can target about 100 genes.

4.4.2 shRNA screenings *in vivo* in human AMLs

One of the main problems we had in accomplishing this task concerned the infection procedure of human AML blasts with the lentiviral library. Most of the human AML blasts derived from our xenograft models did not survive in culture more than 16-24 hours (except for the sample AML-IEO20, that grew indefinitely in liquid culture) and, likely due to the fact that they were suffering, did not get infected. We thus tried many different infection and culture conditions and finally we managed to infect 4 different AML samples: AML5 (t(1;2); NPM wt; FLT3 mutated); AML9 (complex karyotype; NPM WT; FLT3 WT); IEO20 (t(9;11); NPM WT; FLT3 WT); IEO23 (normal karyotype, mutant NPM, mutant FLT3). For AML-IEO20, the only sample able to growth *in vitro*, we also performed a puromycin selection for 96 hours before transplantation. The percentages of infection are listed in Table 16.

Table 16. Percentage of infected cells with the shRNA library in different AML samples. The "incubation time" indicates how long the sample has been kept in culture before being transplanted in recipient animals.

AML	Percentage of infection	Incubation time	Transplanted Mice
AML5	1.20%	13h	2
AML9	21%	3h	2
AML-IEO23	20%	13h	2
AML-IEO20	10%	21h	3

The samples were infected with a lentiviral library custom made by Cellecta Inc. containing 1204 shRNAs targeting 118 genes (10 shRNA for each gene plus positive and negative controls) that code for epigenetic modifiers involved in cancer development.

Before the transplantation, half of the samples was kept as reference and the other half transplanted in 2-3 recipient mice. As soon as the transplanted animals became sick we sacrificed them, collected the human blasts, extracted the DNA along with their reference, and used it to construct the library for high throughput sequencing.

4.4.3 Library preparation

Mice transplanted with infected AML blasts were monitored weekly by peripheral blood FACS analysis. As soon as the level of engraftment in the recipient mice reached 70% to 90%, the animals were euthanized and the blasts from bone marrow and spleen were collected for DNA extraction and library construction. In Figure 34 is shown an example of library preparation obtained at the end of the amplification process for the DNA samples derived from animals transplanted with infected AML5. As shown in Table 16, although this is the sample with the lowest percentage of infection, nethertheless we were able to get the library for the sequencing step. In all the shRNA screening experiments, we could produce the library for sequencing also from the *in vitro* post infection reference DNA, probably thanks to the larger amount of starting DNA (25 million cells) compared to the clonal tracking experiments (half million cells).

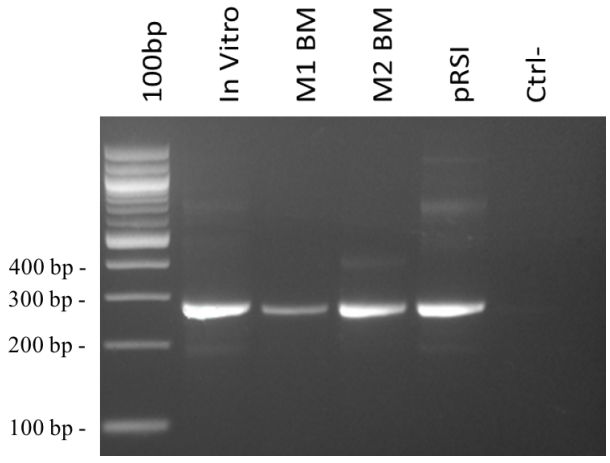

Figure 34. Library preparation from AML5 derived DNA samples. The PCR products were analyzed by 3,5% agarose gel. The band has the expected size of 259 bp. 100bp: 100bp marker; M1 and M2: DNA samples extracted from the bone marrow of two mice transplanted with the AML5 infected blasts; Ctrl: negative control (no DNA); pRSI: positive control (purified plasmid).

4.4.4 Barcodes analysis

The sequencing reads were aligned with "perfect match" to the library sequences provided by Cellecta in order to compare the frequency of each shRNA in the original population soon after the infection (referred as "*in vitro*"), and their frequency after expansion of the tumor *in vivo*.

We first analyzed the reference samples. In these samples all the shRNAs are represented with a similar frequency (Figure 35) calculated as:

$$shRNA\ frequency = \frac{Number\ of\ reads\ for\ a\ specific\ shRNA}{Total\ number\ of\ reads}$$

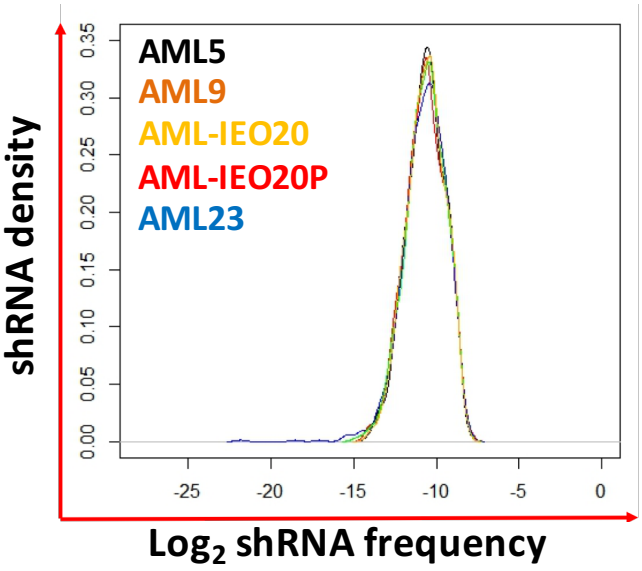

Figure 35. *In vitro* distribution of the 1204 shRNAs. On the *x* axis are reported the $\log_2$ of the shRNAs frequencies, while on *y* axis the shRNA densities. Each line represents an "*in vitro*" AML sample.

These data confirmed that the infection took place in a uniform manner and that all the shRNA are equally represented in the starting population, moreover, it was a good indication that the process of construction of the library was homogeneous and did not introduce main artefacts. We have also considered for each shRNA the second barcode, since, as described in detailed in section 3.7.2, the library is constructed in a way that each

shRNA is associated with up to 12,415 different barcodes, allowing a "clonal" analysis of each shRNAs reads. In Table 17 are shown for each sample "*in vitro*" the number of reads and the number of "clones". According to these numbers, each shRNA is associated with an average of 370 barcodes. The frequency distribution of different barcodes for each shRNA is comparable (not shown).

Table 17. Total number of reads and clones in the "*in vitro*" samples.

Sample	# reads	# clones
AML5	1.49E+07	3.70E+05
AML9	2.23E+07	2.55E+05
AML-IEO20	5.00E+07	4.13E+05
AML-IEO20p	2.66E+07	5.20E+05
AML23	4.08E+08	2.90E+05

We then analyzed the samples after the *in vivo* growth of the tumor. We immediately noticed that all samples presented 2-3 highly expanded shRNAs that alone made up a large part of the whole sample. This expansion was more evident in some animals as: AML9_M1, AML9_M2, AML23_M1, AML-IEO20P_M2, AML-IEO20P_M3 (Figure 36).

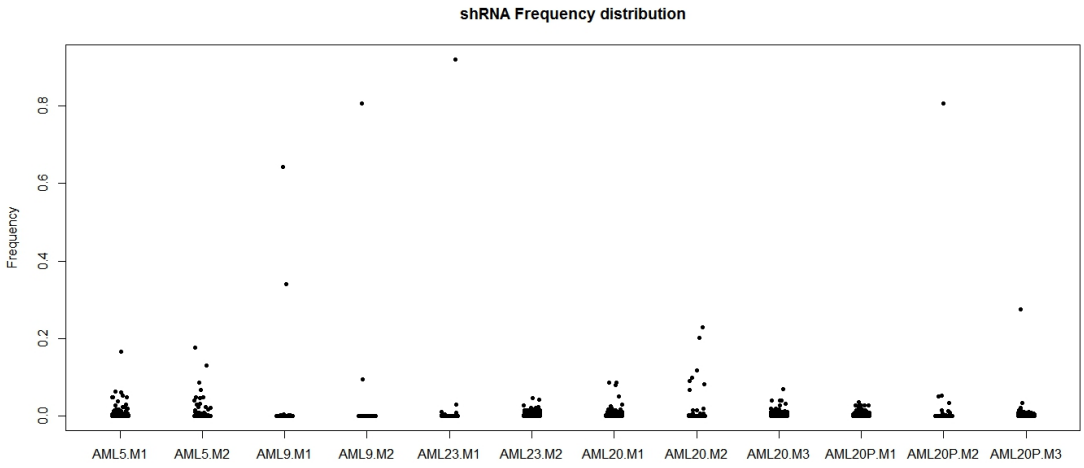

Figure 36. shRNA frequencies distribution. On the *x* axis are represented all the samples derived from a single mouse after *in vivo* growth. On the *y* axis is reported the relative frequency of each shRNA in the sample.

We argued that it was unlikely due to the biological effect of the shRNA, mainly for two reasons: i) the expanded shRNA was different in each mouse (even when transplanted with the same leukemia), and ii) it was not confirmed by the other 9 shRNA specific for the same gene. Our current hypothesis is that a single AML clone, derived from a single cell with a strong growth advantage, grew up *in vivo* generating a large part of the tumor. In support to this interpretation, when we examined all the barcodes associated with the expanded shRNA sequence we found that 99% of the reads were associated to a single barcode thus confirming the "clonal expansion" (Table 18).

Table 18. Example of clonal expansion in the AML-IEO23 sample. 129 clones have been assigned to a SMARCB1 shRNA. The first clone represents the 99.88 % of the reads of the entire population for this shRNA. Highlighted in red is the clone eliminated after applying the 99.5%ile filter.

shRNA	Barcode CT	# reads	n clones
SMARCB1.553	**barfir_3814**	**740603**	**x1**
SMARCB1.553	barfir_9452	150	x1
SMARCB1.553	barfir_3806	49	x1
SMARCB1.553	barfir_3807	45	x1
SMARCB1.553	barfir_3811	44	x1
SMARCB1.553	barfir_9821	43	x1
SMARCB1.553	barfir_9669	29	x1
SMARCB1.553	barfir_9509	26	x1
SMARCB1.553	barfir_3810	22	x1
SMARCB1.553	barfir_7961	21	x1
SMARCB1.553	barfir_7965	20	x1
SMARCB1.553	barfir_#####	15	x2
SMARCB1.553	barfir_#####	14	x2
SMARCB1.553	barfir_#####	13	x3
SMARCB1.553	barfir_#####	11	x4
SMARCB1.553	barfir_#####	10	x3
SMARCB1.553	barfir_#####	9	x2
SMARCB1.553	barfir_#####	8	x2
SMARCB1.553	barfir_#####	7	x9
SMARCB1.553	barfir_#####	6	x4
SMARCB1.553	barfir_#####	5	x5
SMARCB1.553	barfir_#####	4	x4
SMARCB1.553	barfir_#####	3	x8
SMARCB1.553	barfir_#####	2	x20
SMARCB1.553	barfir_#####	1	x47
Total number of reads		741514	
Reads analysed		**911 (0,12%)**	

The clonal expansion of few clones prevented the correct comparison of the frequency of different shRNAs before and after *in vivo* growth. In order to correct our data for the "clonality" issue, we have applied two alternative corrective methods:

1. For each shRNA we have considered the reads up to the 99,5th percentile of the distribution of the second barcodes therefore eliminating from the analysis the outlier clones (see Table 18, as example).

2. We have standardized the raw data applying the Z-score statistical method that uses the mean and standard deviation to transform raw scores in standard scores.

Method 1:

As shown in Table 19, this filter strongly reduced the number of reads considered for the analysis, therefore, after filtering, we have re-calculated the frequency of each shRNA within the population and compared their distribution to the ones obtained in the "*in vitro*" sample.

Table 19. Number of reads and clones considered in shRNA screening. For each animal is reported the total number of reads obtained by NGS (# reads), the total number of clones (# clones) and the percentage of reads used for the analysis after applying the 99,5th percentile filter.

Sample	# reads	# clones	% reads
AML5 M1	1.67E+07	1.68E+04	32.99
AML5 M2	2.80E+07	3.07E+04	8.76
AML9 M1	1.75E+07	7.27E+03	1.8
AML9 M2	2.23E+07	6.55E+04	9.89
AML-IEO20 M1	4.56E+07	4.53E+04	20.81
AML-IEO20 M2	1.86E+07	1.54E+04	5.53
AML-IEO20 M3	3.25E+07	7.42E+04	30.22
AML-IEO20P M1	2.21E+07	5.04E+04	54.45
AML-IEO20P M2	2.29E+07	3.29E+03	10.97
AML-IEO20P M3	4.52E+07	9.67E+04	28.5
AML-IEO23 M1	2.45E+07	6.79E+03	4.06
AML-IEO23 M2	6.02E+07	1.53E+05	41.27

As shown in Figure 37, the shRNAs frequencies distribution after *in vivo* growth is quite variable among mice and is characterized by a large number of depleted shRNAs.

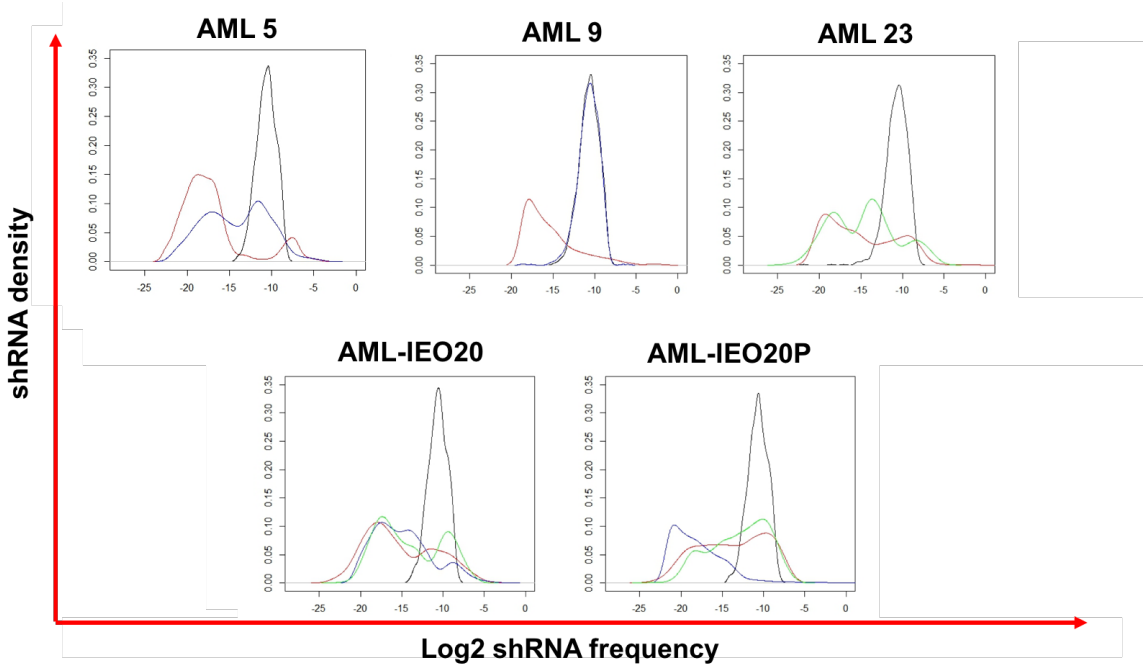

Figure 37. Distribution curves of the shRNAs frequencies *in vivo* (colored) and *in vitro* (black). Each color represents a single mouse.

For comparative analysis, we have calculated the reads average for the 10 shRNA targeting the same gene, the relative frequency in the sample and then the enrichment or depletion compared to the reference with the following formula (FC Fold Change):

$$\log_2 FC = \log_2 \frac{in\ vivo\ frequency}{in\ vitro\ frequency}$$

Method 2:

We have calculated the frequency for each shRNA within the *in vitro* and the *in vivo* AML populations. Then, we have calculated the log2FC for each shRNA with the same formula reported above and, for each log2FC value, we have calculated the Z-score according to the following formula:

$$Z - score = \frac{log2FC - average\ of\ log2FCs}{standard\ deviation\ of\ log2FC}$$

In Figure 38 are shown the z-score distributions of the shRNAs log2FC for each animal.

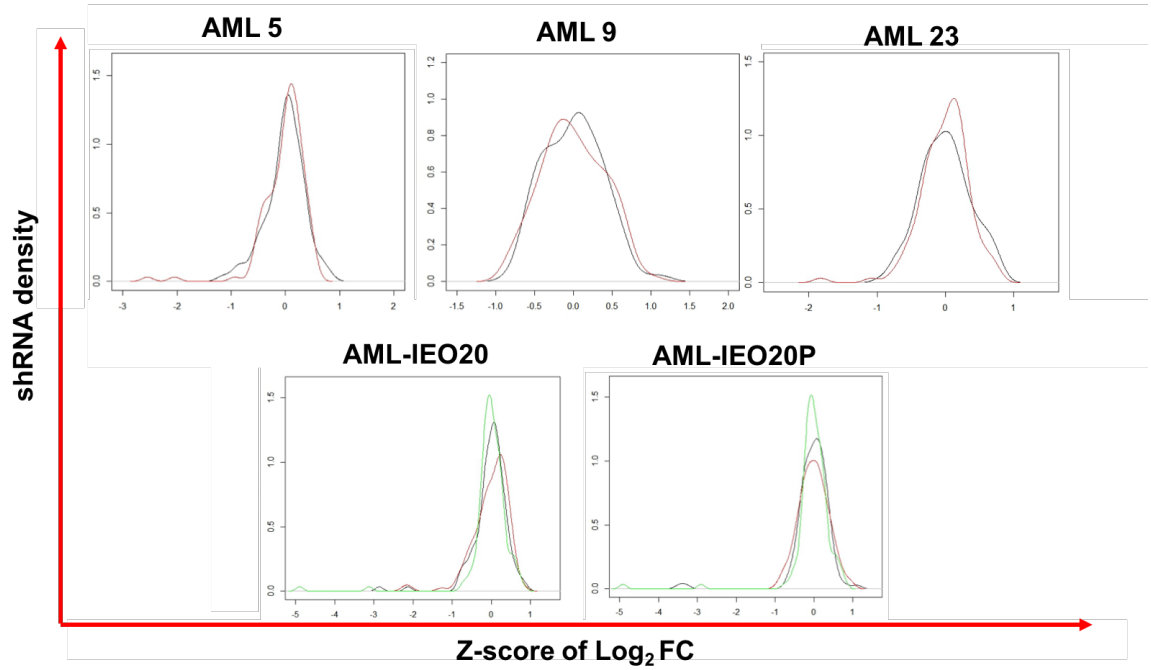

Figure 38. z-score distributions of the shRNAs log2FC for each animal. Each color represents a single mouse.

For comparative analysis, we have calculated the average among the 10 Z-score-log2FC values targeting the same gene.

At this point, we asked how the FC values for each mouse of a given AML sample correlate each other in the two methods. As shown in Table 20, the Pearson coefficients among animals are low, suggesting that the variability of the system is quite high, but are generally increased upon the introduction of the two normalizing methods. In particular, the 99.5%ile filter seems to work better with the AML9 and AML23 samples where the clonality issue was more dramatic, while the Z-score normalization is more effective in the others.

Table 20. Pearson correlations of the log2FC values among different animals. The mice were transplanted with the same leukaemia considering the raw, the 99,5th%ile and the z-score data set.

Total				99,5%ile				Z SCORE			
IEO20	M1	M2	M3	IEO20	M1	M2	M3	IEO20	M1	M2	M3
M1		-0,029	0,099	M1		0,061	0,287	M1		0,57	0,685
M2			0,007	M2			0,013	M2			0,527
M3				M3				M3			
IEO20 Puro	M1	M2	M3	IEO20 Puro	M1	M2	M3	IEO20 Puro	M1	M2	M3
M1		0,1	0,355	M1		-0,08	0,509	M1		0,193	0,76
M2			0,002	M2			-0,12	M2			0,204
M3				M3				M3			
AML5	-0,022			AML5	0,085			AML5	0,249		
AML9	-0,012			AML9	0,213			AML9	-0,035		
AML23	-0,05			AML23	0,189			AML23	0,099		

We concentrated our attention on depleted shRNAs since they should target putative oncogenes that, in turn, can identify new therapeutic targets. In particular, we have ranked the FC values for each mouse and we have considered the shRNA targeted genes that fell in the first 15th%ile. The number of genes that are commonly targeted in more than one animal in the two methods are shown in Table 21.

Table 21. Numbers of common targets among different animals. Dataset comparison between 99,5 and z-score analysis.

Number of Common Targets					
FC	AML5	AML9	AML23	IEO20	IEO20P
99,5	3/19	14/19	2/19	10/19	10/19
Z-score	5/19	3/19	10/19	9/19	15/19

Since in our experimental setting each AML has been divided in 2/3 animals and the "clonality" issue may have partially compromised our ability to cover library complexity within each single animal, we have decided, for each AML, to consider as depleted all the targeted genes resulting from the union of the lists coming from each animal. At this point, we have looked how many genes we have in common for each AML between the 99.5%ile

and the Z-score list. As shown in Figure 39, many genes are in common between the two methods.

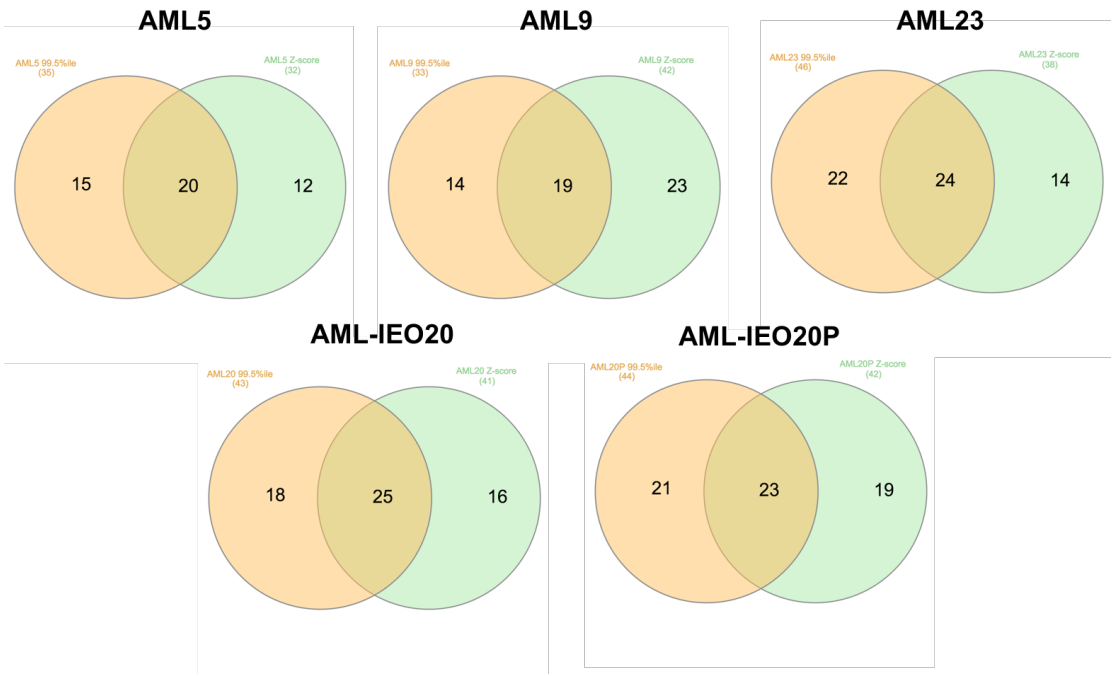

Figure 39. Common target for each AML between 99.5%ile and the Z-score lists. In orange is represented the 99.5%ile list, in green the Z-score list.

Finally, for each AML we have considered only the genes targeted by depleted shRNAs in both the analysis. In Table 22, we show the genes both specific or in common to more than one AML. Strikingly, the two positive controls (PSMA1 and RPL30) are common to many samples, while the negative control (luciferase) is absent. In bold, the genes we have selected to be validated *in vivo*. The validation strategy is described in the "discussion" section.

Table 22. Genes targeted by depleted shRNA in different AML. In red the positive control. In bold the gene we have selected for the validation step.

AMLs					Total in common	gene
AML20	AML20P	AML23	AML5	AML9	1	**PSMA1**
AML20	AML20P	AML5	AML9		1	**BRD9**
AML20	AML20P	AML23	AML5		2	**TRIM66, RPL30**
AML20	AML5	AML9			2	UBE2A, **TRIM24**
AML23	AML5	AML9			1	**CBX5**
AML20	AML20P	AML9			1	PRMT7
AML20	AML23	AML9			2	**PRDM6**, SETD7
AML20P	AML23	AML9			1	**SETD1B**
AML5	AML9				1	BRD7
AML20	AML5				3	CHD9, SETDB2, KDM3A
AML23	AML5				2	UBE2I, CBX6
AML23	AML9				1	INO80
AML20	AML20P				4	SMC3, **PRDM9, JMJD6, KDM4C**
AML20	AML23				3	TET1, ING3, SETD4
AML20P	AML23				1	DNMT3L
AML5					6	SMARCB1, CHD3, SMARCA2, PARPA1, KAT2A, KDM2A
AML9					8	**PRMT1**, DNMT3A, TET3, CHD1L, CHD4, KDM8. PRMT8, TET2
AML20					5	SETD1A, KDM4D, SMARCA4, SMC5, CBX2
AML20P					5	CBX4, MTA3, CHD2, PRDM13, KDM1B
AML23					6	PRDM16, KDM5D, SMC2, CTCF, PRDM8

5 DISCUSSION

The ability of cancer to evolve and adapt is a major challenge for therapy. Genetic plasticity, in fact, is one of the most important characteristic of cancer and consists in the sequential acquisition of somatic mutations that are selected by the tumor in order to acquire multiple cancer hallmarks[26]. Since the pool of somatic mutations is cell specific, within a cancer can coexist multiple genetically and functionally heterogeneous subpopulations. For these reasons, and given the increasing evidence of the heterogeneity of the cancer stem cells, determining the clonal composition inside the tumor population and the behavior of each cancer cell during tumor evolution is essential for understanding cancer biology and driving the study of new, more efficient, therapeutic approaches.

The development of new technologies, that couple high-throughput sequencing to viral genetic barcoding, has allowed to obtain high resolution descriptions of *in vivo* cells dynamics with particular interest to clonality. In particular, using lentiviral-based approaches, a clonal tag (barcode) is stably incorporated into the genomic DNA so that each cell of a clone is identified by a unique barcode that can be resolved through sequence-based detection[100]. These technologies significantly increase the sensitivity of the analysis compared to PCR-based integration site identification, mainly biased by the choice of the restriction enzymes and by intrinsic bias of PCR amplification. The major benefits of this experimental approach are that it is unbiased, quantitative and provides accurate information on the relative proportion of different clones[100].

Up to date this approach has been largely used to study the clonal evolution of normal human and murine hematopoietic stem cells (HSCs) isolated from cordon blood or from bone marrow[100-103]. In particular, these studies addressed the issue of lineage differentiation during hematopoiesis, following the fate of barcoded HSCs. The authors have shown that, both in mouse and human, most cellular lineages derive from a small number of HSCs, however, not all HSCs have the same potential to give rise to more mature populations. Moreover, it has been observed that the contribution of different

clones dynamically changes with time. Indeed, there are directional changes (expansion/decline) and age-dependencies in the clonal output of individual HSCs.

In our study we wanted to apply the same approach to track leukemia initiating cells (LICs) and follow their growth *in vivo* in order to understand how they contribute to tumor development and how they evolve. As experimental system we have used two leukemias: a murine leukemia derived from the PML-RARa knock-in model[43] and a human leukemia engrafted in a immuonocompromised mouse model[75] that bears the t(9;11) translocation and, therefore, expresses the MLL-AF9 fusion protein. This human sample is particularly aggressive compared to other human leukemia xenografts we have developed and allowed us to perform few rounds of transplantation in a relatively short time frame. As described in the "Results" section, for both leukemias, we have decided to analyze the data obtained from our clonal tracking experiments both from a “horizontal” and “vertical” perspective. The “horizontal” perspective allowed us to compare how the leukemic clones behave in different animals transplanted with the same leukemic sample within the same passage of transplantation. The “vertical” perspective allowed us to determine how the leukemic clones evolve in serial passages of transplantation.

One of the main problems we faced in interpreting our data was related to the fact that most of the barcodes identified are represented by a very low number of sequencing reads and their contribution to the tumor mass is extremely low. These sequences could reflect the presence of barcoded cells that were transplanted into the recipient mice but were unable to proliferate or they could be due to background noise sequences, as previously shown[104]. In order to correct for technical bias and to determine the minimum number of cells detectable in our experimental system, we employed the use of a spike-in control, that had been previously applied in experiments of clonal tracking via barcoding in the mammary system[105]. The usage of "spike-in" single barcodes at known frequency before library preparation allowed us to understand that our system has a detection limit within 10 to 100 cells. Moreover, it provided a formal proof that in our experimental settings, we

have a linear correlation between the number of sequence reads and the number of cells present in the samples. Interestingly, most of the low frequency barcodes are represented by a number of reads that are below our experimental detection limit, suggesting that they likely represents the background of our experimental system. We can not formally exclude that they represents randomly selected single cells that did not cycle (putative progenitors), however they account for cellular population that are represented by less than 10 cells and are never detectable in subsequent passages of transplantation, therefore, these clones do not contribute to tumor evolution.

The clonal analysis at the first passage (X1) of transplantation provides the unique opportunity to evaluate the number of LICs present in the original leukemic population. Up to date, the gold standard method to determine the LIC frequency is the limiting dilution cell transplantation assay, in which leukemic cells are transplanted into recipient mice at decreasing doses; the proportion of animals that develop the disease is used to calculate, with a confidence interval, the LIC frequency in the original tumor sample. The main limits of this approach are: i) requires a large number of animals in order to reach statistical significance; ii) single LIC transplantation may be too stressful and alter the results. Based on the considerations listed above, for all our clonal analysis we decided to consider as real clones only the one with a number of reads above the threshold that corresponds to 10 cells. Strikingly, this cut-off allowed us to cover more than 99% of the tumor in all samples analysed. Moreover, the number of LICs we could estimate is very close to the number obtained by limiting dilution transplantation assays. These results suggest that, through our experimental strategy, we identified *bona fide* LICs and that the limiting dilution conditions do not seem to interfere with the ability of the LIC to engraft and grow *in vivo*. However, we do not know if with our approach of transplantation "in bulk" (500,000 leukemic cells/mouse) we are saturating the system. In other words, we cannot exclude that the number of LICs transplanted was higher than the one we report, but we could not score it simply because not all transplanted LICs had the chance to

engraft due to an intrinsic "space" constrain *in vivo*. In order to address this issue, we are currently performing new experiments in which we transplant increasing amount of infected blasts to understand if we are working under limiting conditions. Nonetheless, our experimental approach, compared to the one in limiting dilution, offers the great advantage to gain information regarding the LIC growing potential, because each LIC is univocally and permanently labelled and can be traced during its evolution.

In the normal hematopoietic tissue, it has been reported that there is a clear asymmetric clonal distribution of HSCs in different skeletal compartments, indicating that different HSCs populate different BM niches even long-term[101]. Therefore, we first assessed which is the behavior of the LICs in term of their ability to populate different hematopoietic organs, asking if different leukemic clones are able to engraft and generate the disease in different hematopoietic organs. We compared the clonal composition of the leukemia in the main leukemic organs of the mouse: spleen and BM. In contrast to what observed for the normal HSCs, we did not score any difference in the clonal architecture of the leukemia in the two compartments. This strongly suggests that there is no clonal tropism for the distribution of the disease within the organism and all clones are distributed equally within the hematopoietic system. From a clinical point of view this would imply that a systemic treatment would have the same effect on all leukemic clones independently from their localization within the leukemic tissues.

Analysing our clonal tracking experiments “orizontally”, therefore comparing the clonal evolution of the two leukemias in different mice but in the same passage of transplantation, we clearly showed that the clonal composition is very similar in terms of number of clones, in terms of frequency distributions and even of identity (in the passages were identity could be addressed). The common clones among all animals are indeed few, they are always the same, account for the vast majority of the tumoral population and, strickingly, within different animals they always grow to occupy the same proportion of the tumor population. These parameters are highly reproducible in all transplanted animals, with only

rare exceptions (two animals). This would indicate that each LIC has finite and limited "space" for proliferation.

These properties of the LICs are reflected in a strong clonal selection observed in our "vertical" analysis, from passage to passage of the leukemia. While we can not exclude that from the X2 to the X3 passage of transplantation this selection could be due to a bias towards only the most represented clones in the sampling of the cells to be transplanted, we have shown that from X1 to X2 this is definitely not the case. Indeed, we observe the disappearance of clones that are present at very high frequency in the X1 passage. Moreover, notably, we scored also a class of LICs that expanded from X1 to X2 and then disappeared at the third round of transplantation. Our data indicates that, indeed, most of the LICs are endow with a limited proliferative potential that exhausts in just one or two rounds of transplantation and are not able to sustain leukemogenesis long-term. Indeed, a very minor proportion of clones (8%) is able to constantly expand throughout the passages. Finally, we never observed the appearance of new clones from one passage to the other, except for very few clones in case of the X2 of the murine leukemia mAPL18. However, all these new clones were present at very low frequency (<0.35%). It is worth to speculate that this could suggest the possible existence of LICs that remain dormant/quiescent in the first passage of transplantation and are able to re-enter the cell cycle only following a further transplantation stress. Further studies are necessary to establish if this could be a true hypothesis and to address this issue we are planning to perform clonal tracking experiments also in presence of treatment with 5-fluorouracil (5FU), that induces HSCs proliferation.

A substantial characteristic of stem cells and cancer stem cells, as well, is believed to be their ability for self-renewal without loss of proliferation capacity with each cell division. Therefore, cancer stem cells are putatively immortal. In contrast to these believes, taken all together, our data strongly support a model in which each LIC possesses an intrinsic and highly variable replicative potential. Each LIC appear to be endowed with a determined

capability of proliferation within the tumoral population and, in particular, each leukemia possesses a definite number of LICs able to propagate the disease long-term. These LICs tend to overgrow all other clones to constitute a very large proportion of the tumor. Moreover, based on the results of our clonal tracking experiments, we can speculate that the vast majority of the LICs are not immortal and tend to get old with time, loosing their fitness and their indefinite replicative potential.

Development of cancer requires the accumulation of genetic and epigenetic changes that each confers a selective advantage to the cell at a specific stage of tumorigenesis. These changes include inactivation of onco-suppressors or activation of oncogenes. At the end, the tumor is the result of the complex alteration and intersection of different cellular pathways and, even though we can map these alterations by whole genome sequencing, is quite difficult to predict the epistatic relationship among them and which one is the most critical for tumor growth in a complex environmental system like the one *in vivo*. These reasons may explain why it has been so difficult to find good therapeutic targets that are usually selected in system *in vitro* and have to be validated first in animal model and, then, in humans. Moreover, we have to consider that in the last years it has been shown that in many tumors included AML, only a small subset of cells is able to reconstitute the entire tumor both using human tumors (xenograft models) or mouse tumors (transgenic models) leading to the Cancer Stem Cell (CSC) theory. The idea that CSCs are responsible for driving tumor growth and metastasis has important implications for the treatment of cancer. At present, most cancer therapies are targeted against the actively dividing, bulk population of tumor cells and not necessarily against the CSCs, which may account for resistance to treatments and recurrence of disease, even after prolonged periods of apparent remission. Increasing evidences suggest that in the future a good therapy should combine treatment direct against both the bulk and the CSC tumor populations. The screening we have performed is innovative under several points of view and overcome many of the problems just highlighted. It has been conducted entirely *in vivo* using AMLs xenograft

mouse models and, therefore, it identifies targets involved in the growth of the primary tumor in its natural environment and, since tumor growth *in vivo* relaies mainly on proliferation and expansion of CSCs, this system, for the first time, is highly selective for putative therapeutic target specific for the CSC population.

Recent studies of the spectrum of somatic genetic alterations in acute myeloid leukemia have identified frequent somatic mutations in genes that encode proteins important in the epigenetic regulation of gene transcription. Genes such as ASXL1, DNMT3A, EZH2, IDH1/2, MLL1 and TET2 have been shown to be frequently mutated and/or involved in chromosomal rearrangements in patients with myeloid malignancies[106]. Mutations in these genes are known, or suspected, to affect the functional roles of these proteins, involved in modifying cytosine nucleotides on DNA and/or altering the state of histone modifications[107-110]. Since epigenetic changes induced by oncogenic stimuli are potentially reversible, the proteins involved in epigenetic regulation are potential druggable candidates. Therefore, we decided to use an shRNA lentiviral library able to interfere with the expression of genes that encode for epigenetic regulators. A further advantage of our screening has been the usage of a small number of mice and blasts allowing conducting the screening in more than one AML and leading to the identification both of tumor-specific and tumor-wide putative targets that might offer solutions of therapeutic intervention in a wide group of AML patients.

Particularly important for this part of the project is to obtain an infected population that covers the complexity of library with only one integration per cells. This is critical because we want to be sure that the biological effect that we observe after screening, is the direct result of the action of one single shRNA. For these reasons, and also according with previous published data[96], we decided to obtain a maximum efficacy infection of 20% and to transplant the cells right after the infection. With these condition we were able to obtain infected populations that cover all the library complexity, and more importantly, with one integration per cell.

During the analysis, the main problem we have encountered is the clonality issue. In some animals the impact has been dramatic (AML5_M2; AML9_M1; AML9_M2, IEO20_M2; IEO23_M1; see figure 36 and table 18), in others the percentage of the tumour not covered by the expanded clones was larger allowing a more reliable analysis of the results. However, it is likely that this problem may have reduced our ability to cover the complexity of our library *in vivo*. Indeed, the correlation of the results obtained by different animals transplanted with the same AML is modest and tend to be lower for animals in which the clonal expansion was higher. For this reason, we considered as putative target genes the union and not the intersection among the lists coming from different animals transplanted with the same AML. Based on these considerations, we think that for the next experiments we need to modify our protocol as following:

- Perform the clonal tracking experiments in order to evaluate how each AML sample grows, both in the xenograft models for human or in the syngeneic recipients for mouse.
- Based on the clonal tracking information, calculate how many infections and animals have to be transplanted in order to have statistically significant data.

For the time being, to face this problem we have applied two different normalizing methods: i) we have eliminated outliers clones from the *in vivo* samples and we have recalculated the shRNA frequencies in the resized samples; ii) we have considered the entire population and we have normalized the fold change values (FC) between the *in vitro* and the *in vivo* samples. Although the two methods are conceptually quite different, the genes targeted by the down-regulated shRNAs are fairly consistent with the two methods, and this prompted us to consider these common genes as putative targets and proceed with the validation protocol.

It is interesting to note that the deregulation and/or overexpression of most of the identified potential targets listed in Table 21 are already reported to be involved in different types of cancer including leukemia. The criteria we have used for the selection of the genes to be

validated are the following: i) recurrence among different AML samples; ii) data available in the literature about their involvement in cancer progression; iii) recognised as possible therapeutic targets to develop specific inhibitors (http://www.thesgc.org/chemical-probes/epigenetics).

The TRIM family proteins are involved in tumorigenesis and tumor progression, affecting pathways such as cell proliferation, DNA repair and apoptosis as for example in promyelocytic leukemia[111]. CBX5 has been shown to be important in the regulation of pathways that control stemness in lung cancer[112]. Interestingly, we found different component of the SWI/SNF chromatin remodelling complexes previously associated with the development of solid tumors as the ovarian[113], liver cancer[114], renal carcinomas[115] and pancreatic cancer[116]. Morever, very recently, it has been reported that SMARCA4 and SMARCD2 are required in the maintenance of AML in a murine model of MLL-AF9 leukemia and, in particular, for leukemia self-renewal[117]. The bromodomain-containg protein 9 (BRD9) is also present in some SWI/SNF ATPase remodelling complexes[118-120] and it has been identified to have a copy number variation (CNV) in samples of non-small cell lung cancer (NSCLC)[121] and in cervical cancer[119]. Moreover, it can be selectively target by a chemical probe[122]. PRDM protein family members are involved in gene expression regulation, modifying the chromatin structure directly, through the intrinsic methyltransferase activity, or indirectly through the recruitment of chromatin remodelling complexes. In our screening we have identified PRDM6, PRDM9, PRDM13 and PRDM16 as potential leukemic target. In particular, PRDM9 is involved in meiotic recombination events and could induce pathological genomic rearrangement[123] and PRDM16 is involved in human leukemic translocations and is highly expressed in some AML[124].

We found other enzymes characterized by methyltransferase activity as SETD1B and the PRMT1. Deregulation of SETD1B is associated with lung cancer[125] and is recurrently amplified in melanoma, accelerating disease onset[126] while overexpression of PRMT1 is associated to breast and prostate cancers[127].

Another interesting class of enzyme we have scored in the screening are the histone lysine demethylases. Deregulation of these proteins has been associated in the development of cancer[128] and inactivating mutations are present in multiple myeloma, carcinomas, glioblastoma, breast and colorectal cancer[129,130]. In particular, KDM4C is deregulated in the aggressive basal breast cancer subtype[131]. JMJD6, as for KDM4C, is involved in breast cancer, its deregulation drives cell proliferation and motility and it is a marker of poor prognosis[132]. Interestingly, JMJD6 was recently reported to interact and required for the activity of BRD4, a widely studied therapeutically target in AML[133].

For the validation step, our main limit is the impossibility to select a homogeneous AML population expressing the shRNA, since, except for AML-IEO20, our AML samples do not grow *in vitro*. To overcome this problem, we have cloned two shRNA for each gene we want to validate in a lentiviral vector in which the expression of the shRNA may be induced by administration of doxycycline, moreover, the same vector constitutively expresses the reporter gene eGFP. Briefly, we will infect and immediately transplant our AML samples with the shRNA expressing lentiviral particles. We will wait the expansion *in vivo* of the mixed (GFP+ and GFP-) AML population, then, we will sacrifice the animal, purify the GFP+ blasts by FACS sorting and we will retransplant them in recipient animals. At this point, we will induce the expression of the specific shRNA by doxycycline administration and we will monitor leukemia development compared to control animals.

At the end of this validation step, we will be able to establish to which extent our screening results are reliable in identifying new putative therapeutic target for AML treatment. The big advantage is that these targets will be already validated in the pre-clinical model.

If, as we believe, a fairly good number of targets will be validated *in vivo*, we can extend this approach to other samples in an effort to correlate classes of targets to molecularly defined AMLs.

At the end of this project, we are confident we will have set up an innovative experimental protocol that can change the way we consider the importance of AML-specific molecular

alterations. So far they have been used mainly for studies on prognosis and risk stratification but their impact on therapeutic choices is poor. These studies will sign a great advance in the possibility, in the future, of having differentiated therapeutic approach specific for each molecular characterized AML subtype that will hopefully increase patients survival and decrease therapy related side effects.

6 REFERENCES

1. Weissman, I. L. Stem Cells: Units of Development, Review Units of Regeneration, and Units in Evolution. *Cell* 157–168 (1999).
2. Weissman, I. L. The road ended up at stem cells. *Immunological review* 159–174 (2002).
3. Notta, F. *et al.* Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment. *science* 218–221 (2015). doi: 10.5061/dryad.c0q0h
4. Seita, J. & Weissman, I. L. Hematopoietic stem cell: self-renewal versus differentiation. *WIREs Syst Biol Med* **2,** 640–653 (2010).
5. Larsson, J. & Karlsson, S. The role of Smad signaling in hematopoiesis. *Oncogene* **24,** 5676–5692 (2005).
6. Rieger, M. A. & Schroeder, T. Hematopoiesis. *Cold Spring Harbor Perspectives in Biology* **4,** a008250–a008250 (2012).
7. Walter, D. *et al.* Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. *Nature* **520,** 549–552 (2015).
8. Pang, W. W. *et al.* Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. *Proceedings of the National Academy of Sciences* **108,** 20012–20017 (2011).
9. Trumpp, A., Essers, M. & Wilson, A. Awakening dormant haematopoietic stem cells. *Nature Reviews Immunology* **10,** 201–209 (2010).
10. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat Immunol* **5,** 738–743 (2004).
11. Huntly, B. J. P. & Gilliland, D. G. Leukaemia stem cells and the evolution of cancer-stem-cell research. *Nat Rev Cancer* 311–322 (2005).
12. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* 1–7 (2001).
13. Tabe, Y. & Konopleva, M. Advances in understanding the leukaemia microenvironment. *Br J Haematol* **164,** 767–778 (2014).
14. Wright, D. E., Wagers, A. J., Gulati, A. P., Johnson, F. L. & Weissman, I. L. Physiological Migration of Hematopoietic Stem and Progenitor Cells. *science* **294,** 1933–1936 (2001).
15. So, C. W. *et al.* MLL-GAS7 transforms multipotent hematopoietic progenitors and induces mixed lineage leukemias in mice. *Cancer Cell* **3,** 161–171 (2003).
16. Martinez-Climent, J. A., Fontan, L., Gascoyne, R. D., Siebert, R. & Prosper, F. Lymphoma stem cells: enough evidence to support their existence? *Haematologica* **95,** 293–302 (2010).
17. Tefferi, A. & Vardiman, J. W. Classification and diagnosis of myeloproliferative neoplasms: The 2008 World Health Organization criteria and point-of-care diagnostic algorithms. *Leukemia* **22,** 14–22 (2007).
18. Vardiman, J. W. *et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114,** 937–951 (2009).
19. Fey, M., Dreyling, M.On behalf of the ESMO Guidelines Working Group. Acute myeloblastic leukemia in adult patients: ESMO Clinical Recommendations for diagnosis, treatment and follow-up. *Annals of Oncology* **20,** iv100–iv101 (2009).
20. Union for International Cancer Control. *ACUTE MYELOGENOUS LEUKEMIA AND ACUTE PROMYELOCYTIC LEUKEMIA . WHO* 1–14 (2014).
21. Bacher, U. *et al.* Comparison of genetic and clinical aspects in patients with acute myeloid leukemia and myelodysplastic syndromes all with more than 50% of

bone marrow erythropoietic cells. *Haematologica* **96,** 1284–1292 (2011).
22. Gregory, T. K. *et al.* Molecular prognostic markers for adult acute myeloid leukemia with normal cytogenetics. *Journal of Hematology & Oncology* **2,** 23–10 (2009).
23. Dohner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* 453–474 (2009). doi:10.1182/blood-2009-07
24. Yoshida, H. *et al.* Accelerated Degradation of PML-Retinoic Acid Receptor a (PML-RARA)Oncoprotein by AH-iraws-Retinoic Acid in Acute Promyelocytic Leukemia: Possible Role of the Proteasome Pathway1. *Cancer Research* 2945–2948 (1996).
25. Gilliland, D. G. The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100,** 1532–1542 (2002).
26. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674 (2011).
27. Takahashi, S. Current findings for recurring mutations in acute myeloid leukemia. *Journal of Hematology & Oncology* **4,** 36 (2011).
28. Frohling, S. Genetics of Myeloid Malignancies: Pathogenetic and Clinical Implications. *Journal of Clinical Oncology* **23,** 6285–6295 (2005).
29. Kelly, L. M. & Gilliland, D. G. GENETICS OF MYELOID LEUKEMIAS. *Annu. Rev. Genom. Human Genet.* **3,** 179–198 (2002).
30. Speck, N. A. & Gilliland, D. G. Core-binding factors in haematopoiesis and leukaemia. *Nat Rev Cancer* **2,** 502–513 (2002).
31. Richters, A. *et al.* Targeting Gain of Function and Resistance Mutations in Abl and KIT by Hybrid Compound Design. *J. Med. Chem.* **56,** 5757–5772 (2013).
32. Kralovics, R. *et al.* A Gain-of-Function Mutation of JAK2in Myeloproliferative Disorders. *N Engl J Med* **352,** 1779–1790 (2005).
33. Krampert, M., Heldin, C.-H. & Heuchel, R. L. A gain-of-function mutation in the PDGFR-β alters the kinetics of injury response in liver and skin. *Lab Invest* **88,** 1204–1214 (2008).
34. Letard, S. *et al.* Gain-of-Function Mutations in the Extracellular Domain of KIT Are Common in Canine Mast Cell Tumors. *Molecular Cancer Research* **6,** 1137–1145 (2008).
35. Dai, C. *et al.* Loss of tumor suppressor NF1 activates HSF1 to promote carcinogenesis. *J. Clin. Invest.* **122,** 3742–3754 (2012).
36. Bowen, M. E. *et al.* Loss-of-Function Mutations in PTPN11 Cause Metachondromatosis, but Not Ollier Disease or Maffucci Syndrome. *PLoS Genet* **7,** e1002050–11 (2011).
37. Renneville, A. *et al.* Cooperating gene mutations in acute myeloid leukemia: a review of the literature. *Leukemia* **22,** 915–931 (2008).
38. Christiansen, D. H., Andersen, M. K., Desta, F. & Pedersen-Bjergaard, J. Mutations of genes in the receptor tyrosine kinase (RTK)/RAS-BRAF signal transduction pathway in therapy-related myelodysplasia and acute myeloid leukemia. *Leukemia* **19,** 2232–2240 (2005).
39. Chen, J., Odenike, O. & Rowley, J. D. Leukaemogenesis: more than mutant genes. *Nat Rev Cancer* **10,** 23–36 (2010).
40. Jan, M. & Majeti, R. Clonal evolution of acute leukemia genomes. **32,** 135–140 (2012).
41. Bueno, C., Montes, R., Catalina, P., guez, R. R. I. & Menendez, P. Insights into the cellular origin and etiology of the infant pro-B acute lymphoblastic leukemia with MLL-AF4 rearrangement. *Leukemia* **25,** 400–410 (2010).
42. Kelly, L. M. *et al.* PML/RAR and FLT3-ITD induce an APL-like disease in a mouse model. *Proceedings of the National Academy of Sciences* **99,** 8283–

8288 (2002).
43. Westervelt, P. High-penetrance mouse model of acute promyelocytic leukemia with very low levels of PML-RAR expression. *Blood* **102,** 1857–1865 (2003).
44. Mallardo, M. *et al.* NPMc+ and FLT3_ITD mutations cooperate in inducing acute leukaemia in a novel mouse model. **27,** 2248–2251 (2013).
45. Schessl, C. *et al.* The AML1-ETO fusion gene and the FLT3 length mutation collaborate in inducing acute leukemia in mice. *J. Clin. Invest.* **115,** 2159–2168 (2005).
46. Stratton, M. R. Exploring the Genomes of Cancer Cells: Progress and Promise. *science* **331,** 1553–1558 (2011).
47. Teixeira, M. R., Pandis, N., Andersen, J. A. & Heim, S. Karyotypic comparisons of multiple tumorous and macroscopically normal surrounding tissue samples from patients with breast cancer. *Cancer Genetics and Cytogenetics* **91,** 124 (1996).
48. Cottu, P. H. *et al.* Intratumoral heterogeneity of HER2/neu expression and its consequences for the management of advanced breast cancer. *Annals of Oncology* **19,** 595–597 (2007).
49. Shipitsin, M. *et al.* Molecular Definition of Breast Tumor Heterogeneity. *Cancer Cell* **11,** 259–273 (2007).
50. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Research* **20,** 68–80 (2010).
51. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
52. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
53. Park, S. Y., Gönen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120,** 636–644 (2010).
54. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109–1113 (2010).
55. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469,** 356–361 (2011).
56. Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366,** 1090–1098 (2012).
57. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481,** 506–509 (2012).
58. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proceedings of the National Academy of Sciences* **105,** 13081–13086 (2008).
59. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N Engl J Med* **366,** 883–892 (2012).
60. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149,** 994–1007 (2012).
61. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).
62. Parkin, B. *et al.* Clonal evolution and devolution after chemotherapy in adult acute myelogenous leukemia. *Blood* **121,** 369–377 (2013).
63. Kronke, J. *et al.* Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. *Blood* **122,** 100–108 (2013).
64. Landau, D. A., Carter, S. L., Getz, G. & Wu, C. J. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia* **28,** 34–43 (2013).
65. Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic

leukemia. *N Engl J Med* **365,** 2497–2506 (2011).
66. Cook, G. J. & Pardee, T. S. Animal models of leukemia: any closer to the real thing? *Cancer Metastasis Rev* **32,** 63–76 (2012).
67. Bruserud, O. *et al.* Effects of interleukin 10 on blast cells derived from patients with acute myelogenous leukemia. *Leukemia* **9,** 1910–1920 (1995).
68. Bruserud, O., Gjertsen, B. T. & Volkman, von, H. L. *In vitro* culture of human acute myelogenous leukemia (AML) cells in serum-free media: studies of native AML blasts and AML cell lines. *J. Hematother. Stem Cell Res.* **9,** 923–932 (2000).
69. Vick, B. *et al.* An Advanced Preclinical Mouse Model for Acute Myeloid Leukemia Using Patients' Cells of Various Genetic Subgroups and *In Vivo* Bioluminescence Imaging. *PLoS ONE* **10,** e0120925–20 (2015).
70. Gould, S. E., Junttila, M. R. & de Sauvage, F. J. Translational value of mouse models in oncology drug development. *Nat Med* **21,** 431–439 (2015).
71. Rangarajan, A. & Weinberg, R. A. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nat Rev Cancer* **3,** 952–959 (2003).
72. Cesano, A. *et al.* The severe combined immunodeficient (SCID) mouse as a model for human myeloid leukemias. *Oncogene* **7,** 827–836 (1992).
73. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367,** 645–648 (1994).
74. Sawyers, C. L., Gishizky, M. L., Quan, S., Golde, D. W. & Witte, O. N. Propagation of human blastic myeloid leukemias in the SCID mouse. *Blood* **79,** 2089–2098 (1992).
75. Agliano, A. *et al.* Human acute leukemia cells injected in NOD/LtSz- scid/IL-2Rγnull mice generate a faster and more efficient disease compared to other NOD/ scid-related strains. *Int. J. Cancer* **123,** 2222–2227 (2008).
76. Ishikawa, F. Development of functional human blood and immune systems in NOD/SCID/IL2 receptor chainnull mice. *Blood* **106,** 1565–1573 (2005).
77. Ishikawa, F. *et al.* Chemotherapy-resistant human AML stem cells home to and engraft within the bone-marrow endosteal region. *Nature Biotechnology* **25,** 1315–1321 (2007).
78. eacute, M. M. *et al.* Stable and reproducible engraftment of primary adult and pediatric acute myeloid leukemia in NSG mice. *Leukemia* **25,** 1635–1639 (2011).
79. Sanchez, P. V. *et al.* A robust xenotransplantation model for acute myeloid leukemia. *Leukemia* **23,** 2109–2117 (2009).
80. Jacoby, E., Chien, C. D. & Fry, T. J. Murine Models of Acute Leukemia: Important Tools in Current Pediatric Leukemia Research. *Frontiers in Oncology* **4,** 1985 (2014).
81. Lock, R. B. *et al.* The nonobese diabetic/severe combined immunodeficient (NOD/SCID) mouse model of childhood acute lymphoblastic leukemia reveals intrinsic differences in biologic characteristics at diagnosis and relapse. *Blood* **99,** 4100–4108 (2002).
82. Cox, C. V. *et al.* Characterization of a progenitor cell population in childhood T-cell acute lymphoblastic leukemia. *Blood* **109,** 674–682 (2007).
83. Shultz, L. D. *et al.* Human Lymphoid and Myeloid Cell Development in NOD/LtSz-scid IL2R null Mice Engrafted with Mobilized Human Hemopoietic Stem Cells. *The Journal of Immunology* **174,** 6477–6489 (2005).
84. Morisot, S. *et al.* High frequencies of leukemia stem cells in poor-outcome childhood precursor-B acute lymphoblastic leukemias. *Leukemia* **24,** 1859–1866 (2010).
85. Wilson, R. C. & Doudna, J. A. Molecular Mechanisms of RNA Interference.

Annu. Rev. Biophys. **42,** 217–239 (2013).
86. Meacham, C. E., Ho, E. E., Dubrovsky, E., Gertler, F. B. & Hemann, M. T. *In vivo* RNAi screening identifies regulators of actin dynamics as key determinants of lymphoma progression. *Nature Genetics* **41,** 1133–1137 (2009).
87. Zender, L. *et al.* An Oncogenomics-Based *In Vivo* RNAi Screen Identifies Tumor Suppressors in Liver Cancer. *Cell* **135,** 852–864 (2008).
88. Beronja, S. *et al.* RNAi screens in mice identify physiological regulators of oncogenic growth. *Nature* **501,** 185–190 (2013).
89. Zuber, J. *et al.* RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478,** 524–528 (2011).
90. Zuber, J. *et al.* An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. *Genes & Development* **25,** 1628–1640 (2011).
91. Kvinlaug, B. T. *et al.* Common and Overlapping Oncogenic Pathways Contribute to the Evolution of Acute Myeloid Leukemias. *Cancer Research* **71,** 4117–4129 (2011).
92. Kurosawa, S. *et al.* Prognostic factors and outcomes of adult patients with acute myeloid leukemia after first relapse. *Haematologica* **95,** 1857–1864 (2010).
93. Hope, K. & Bhatia, M. Clonal interrogation of stem cells. *Nature Publishing Group* **8,** S36–S40 (2011).
94. Shultz, L. D., Ishikawa, F. & Greiner, D. L. Humanized mice in translational biomedical research. *Nature Reviews Immunology* **7,** 118–130 (2007).
95. Simpson-Abelson, M. R. *et al.* Long-Term Engraftment and Expansion of Tumor-Derived Memory T Cells Following the Implantation of Non-Disrupted Pieces of Human Lung Tumor into NOD-scid IL2R null Mice. *The Journal of Immunology* **180,** 7009–7018 (2008).
96. Barde, I., Salmon, P. & Trono, D. *Production and Titration of Lentiviral Vectors*. 1–23 (John Wiley & Sons, Inc., 2001). doi:10.1002/0471142301.ns0421s53
97. Minucci, S. *et al.* PML-RAR induces promyelocytic leukemias with high efficiency following retroviral gene transfer into purified murine hematopoietic progenitors. *Blood* **100,** 2989–2995 (2002).
98. Hu, Y. & Smyth, G. K. ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of Immunological Methods* **347,** 70–78 (2009).
99. Bric, A. *et al.* Functional Identification of Tumor-Suppressor Genes through an *In Vivo* RNA Interference Screen in a Mouse Lymphoma Model. *Cancer Cell* **16,** 324–335 (2009).
100. Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115,** 2610–2618 (2010).
101. Verovskaya, E. *et al.* Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. *Blood* **122,** 523–532 (2013).
102. Cheung, A. M. S. *et al.* Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* **122,** 3129–3137 (2013).
103. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496,** 229–232 (2013).
104. Neff, N. F., Quake, S. R., Lu, R. & Weissman, I. L. Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature Biotechnology* **29,** 928–933 (2011).
105. Nguyen, L. V. *et al.* DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts. *Nature*

Communications **5,** 1–9 (1AD).
106. Chung, Y. R., Schatoff, E. & Abdel-Wahab, O. Epigenetic alterations in hematopoietic malignancies. *Int J Hematol* **96,** 413–427 (2012).
107. Conway O'Brien, E., Prideaux, S. & Chevassut, T. The Epigenetic Landscape of Acute Myeloid Leukemia. *Advances in Hematology* **2014,** 1–15 (2014).
108. Zaidi, S. K. *et al.* Epigenetic mechanisms in leukemia. *Adv Biol Regul* **52,** 369–376 (2012).
109. Liu, J. *et al.* Promyelocytic leukemia protein interacts with werner syndrome helicase and regulates double-strand break repair in γ-irradiation-induced DNA damage responses. *Biochemistry Moscow* **76,** 550–554 (2011).
110. Zhong, X. & Safa, A. R. Phosphorylation of RNA Helicase A by DNA-Dependent Protein Kinase Is Indispensable for Expression of the MDR1Gene Product P-Glycoprotein in Multidrug-Resistant Human Leukemia Cells †. *Biochemistry* **46,** 5766–5775 (2007).
111. Cambiaghi, V. *et al.* TRIM proteins in cancer. *Adv. Exp. Med. Biol.* **770,** 77–91 (2012).
112. Yu, Y.-H. *et al.* Network Biology of Tumor Stem-like Cells Identified a Regulatory Role of CBX5 in Lung Cancer. *Sci. Rep.* **2,** 1–9 (2012).
113. Wiegand, K. C. *et al.* ARID1AMutations in Endometriosis-Associated Ovarian Carcinomas. *N Engl J Med* **363,** 1532–1543 (2010).
114. Li, M. *et al.* Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nature Genetics* **43,** 828–829 (2011).
115. Varela, I. *et al.* Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 1–5 (2011). doi:10.1038/nature09639
116. Shain, A. H. *et al.* Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer. *Proceedings of the National Academy of Sciences* **109,** E252–E259 (2012).
117. Cruickshank, V. A. *et al.* SWI/SNF Subunits SMARCA4, SMARCD2 and DPF2 Collaborate in MLL-Rearranged Leukaemia Maintenance. *PLoS ONE* **10,** e0142806 (2015).
118. Middeljans, E. *et al.* SS18 Together with Animal-Specific Factors Defines Human BAF-Type SWI/SNF Complexes. *PLoS ONE* **7,** e33834–10 (2012).
119. Scotto, L. *et al.* Integrative genomics analysis of chromosome 5p gain in cervical cancer reveals target over-expressed genes, including Drosha. *Mol Cancer* **7,** 58–10 (2008).
120. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nature Genetics* **45,** 592–601 (2013).
121. Kang, J. U., Koo, S. H., Kwon, K. C., Park, J. W. & Kim, J. M. Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer. *Cancer Genetics and Cytogenetics* **182,** 1–11 (2008).
122. Theodoulou, N. H. *et al.* Discovery of I-BRD9, a Selective Cell Active Chemical Probe for Bromodomain Containing Protein 9 Inhibition. *J. Med. Chem.* 150430080108003–15 (2015). doi:10.1021/acs.jmedchem.5b00256
123. Hussin, J. *et al.* Rare allelic forms of PRDM9 associated with childhood leukemogenesis. *Genome Research* **23,** 419–430 (2013).
124. Aguilo, F. *et al.* Prdm16 is a physiologic regulator of hematopoietic stem cells. *Blood* **117,** 5057–5066 (2011).
125. Rodriguez-Paredes, M. *et al.* Gene amplification of the histone methyltransferase SETDB1 contributes to human lung tumorigenesis. **33,** 2807–2813 (2013).

126. Ceol, C. J. *et al.* The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471,** 513–517 (2011).
127. Yang, Y. & Bedford, M. T. Protein arginine methyltransferases and cancer. *Nat Rev Cancer* **13,** 37–50 (2012).
128. Højfeldt, J. W., Agger, K. & Helin, K. Histone lysine demethylases as targets for anticancer therapy. *Nature Publishing Group* **12,** 917–930 (2013).
129. van Haaften, G. *et al.* Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nature Genetics* **41,** 521–523 (2009).
130. Dalgliesh, G. L. *et al.* Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463,** 360–363 (2010).
131. Liu, G. *et al.* Genomic amplification and oncogenic properties of the GASC1 histone demethylase gene in breast cancer. *Oncogene* **28,** 4491–4500 (2009).
132. Lee, Y. F. *et al.* JMJD6 is a driver of cellular proliferation and motility and a marker of poor prognosis in breast cancer. *Breast Cancer Research* **14,** R85 (2012).
133. Liu, W. *et al.* Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release. *Cell* **155,** 1581–1595 (2013).