Whole Exome Sequencing of Chronic Lymphocytic Leukemia (CLL): Extending the Number of Recurrently Mutated Genes in CLL and Detailed Analysis of the Putative CLL Driver Mutations in *XPO1*

Dissertation der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

vorgelegt von

Nikola Konstandin
aus Regensburg

München, 2015

Dissertation eingereicht am: 20.01.2015

Erstgutachter: Prof. Dr. Heinrich Leonhardt

Zweitgutachter: Prof. Dr. Dirk Eick

Tag der mündlichen Prüfung: 07.07.2015

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

München, den ....

(Unterschrift)

Erklärung

Hiermit erkläre ich, *

☐ dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

☐ dass ich mich anderweitig einer Doktorprüfung ohne Erfolg **nicht** unterzogen habe.

☐ dass ich mich mit Erfolg der Doktorprüfung im Hauptfach .....

und in den Nebenfächern .....

bei der Fakultät für ..... der .....

(Hochschule/Universität)

unterzogen habe.

☐ dass ich ohne Erfolg versucht habe, eine Dissertation einzureichen oder mich der Doktorprüfung zu unterziehen.

München, den..... .....

(Unterschrift)

*) Nichtzutreffendes streichen

Meinen Eltern

INDEX

1 Introduction

1.1 Adult Hematopoiesis

Hematopoiesis is derived from the ancient Greek words αἷμα (blood) and ποιεῖν (to make). This term describes the continuous and complex process of blood cell formation. The production of blood cells, which is in adults primarily restricted to the bone marrow, involves cell renewal, proliferation, differentiation and maturation. The basis of this process are hematopoietic growth factors and cytokines but most importantly the hematopoietic stem cells (HSC), a small cell population which is capable of self-renewal and has the ability to give rise to differentiated progeny. Blood cells are mainly subdivided into lymphoid and myeloid-erythroid cells. The lymphoid lineage consists of B-cells, T-cells and natural killer cells. These leukocytes are part of the immune system. B-cells produce highly specific antibodies against antigens and develop into memory B-cells upon activation. There is a variety of T-cells with different functions. They can coordinate immune response or directly attack infected or cancerous cells. Natural killer cells also recognize and eliminate foreign cells. Erythrocytes, which are derived from the myeloid-erythroid lineage, are responsible for the transport of respiratory gases. Thrombocytes, granulocytes and monocytes are the leukocytic fraction of this lineage and they are responsible for coagulation, immune response and phagocytosis (after differentiation), respectively.

In 1961 Till and McCulloch conducted a series of experiments which revealed the ability of bone marrow cells to confer engraftment to lethally irradiated mice (Till and Mc 1961). In 1994 Morrison and Weissman isolated two types of multipotent cell types, the long-term hematopoietic stem cell (LT-HSC) and the short-term hematopoietic stem cell (ST-HSC) (Morrison and Weissman 1994). LT-HSCs are characterized by their ability of self-renewal for life in a lethally irradiated recipient. ST-HSCs originate from LT-HSCs and have only limited self-renewal capacity (Weissman 2000). These, in turn, differentiate into multipotent progenitors (MPPs). They have the ability to give rise to oligolineage progenitors, from which the differentiated progeny is generated. The committed oligolineage progenitors derived from MPPs are the common myeloid progenitors (CMP), which give rise to the myeloid-erythroid lineage, and the common lymphoid progenitors (CLP) generating the lymphoid lineage (B-cells, T-cells and natural killer cells). CMPs then differentiate into megakaryocytic-erythrocyte progenitors (MEP) and granulocyte-monocyte progenitor (GMP) before they differentiate and proliferate into lineage specific mature blood cells: megakaryocytes or erythrocytes and granulocytes, monocytes or mast cells (Iwasaki and Akashi 2007). Dendritic cells can be generated from myeloid and lymphoid progenitors. Mature dendritic cells are found outside the blood and function as antigen-presenting cells.

1.2 Leukemia

Leukemias are hematological disorders, which develop from the clonal proliferation of malignant white blood cells, also known as blast cells that accumulate in bone marrow and blood. Leukemias can be generally divided in to acute and chronic forms. Acute leukemias are usually very aggressive and require immediate treatment due to a rapid increase of immature blood cells. Genetic lesions are thought to cause an increased rate of proliferation, reduced apoptosis and block of cellular differentiation in these cells. Acute leukemias are defined by the presence of a minimum of 20 % blast cells in the bone marrow. Chronic leukemias, in contrast, are characterized by the appearance of more mature but still abnormal cells and longer disease progression therefore an immediate treatment is usually not necessary. Depending on the affected cell type, leukemias are further subdivided in myeloid and lymphoid leukemia. Despite the many different subtypes, leukemias and other hematological malignancies (Non-Hodgin–lymphoma, Myeloma and Hodgkin lymphoma) represent only 7 % of all malignant diseases in females and 9 % in males (Smith *et al.* 2010).

1.2.1 Acute Myeloid Leukemia (AML)

AML has an age-standardized incidence of approximately 3 in 100,000 with a median onset of >60 years (Bishop 1999). It is the most common form of acute leukemia in adults. It is characterized by abnormal blasts in the bone marrow that are also found in the peripheral blood. Initially, the different subtypes of AML were defined based on morphologic parameters, mainly on the degree of maturation. The so called FAB-classification (French-American-British) has been replaced by the classification of the World Health Organization (WHO), which is mainly based on cytogenetic and molecular genetic abnormalities.

1.2.2 Chronic Myeloid Leukemia (CML)

CML is a malignant clonal disorder of hematopoietic stem cells. In addition to myeloid cells, there is also an increase of erythroid cells and platelets in the blood. The annual incidence is between 1-2 cases per 100,000, the median age is 60-65 years (Baccarani and Dreyling 2010). From an initial benign chronic phase the disease turns into a stage of blast crisis within a view years. The final stage of blast crisis, which is similar to an acute leukemia, is often preceded by an accelerated phase. The cytogenetic hallmark of this disease is the reciprocal t(9;22)(q34;q11) chromosomal translocation. This translocation results in the BCR/ABL1 fusion gene, the resulting BCR/ABL fusion protein has tyrosine kinase activity. Therefore patients are treated with tyrosine kinase inhibitors. The progression of disease is frequently accompanied by additional chromosomal abnormalities.

1.2.3 Acute Lymphocytic Leukemia (ALL)

ALL is characterized by the malignant transformation of lymphoid progenitor cells. Most of the ALLs derive from B cells, but they can also derive from T cells. Malignant cells mostly accumulate in bone marrow but also spill into the peripheral circulation. In children, ALL is

much more common than in adults. A number of translocations have been found in this disease. In childhood ALL, the most common translocation is the t(12;21) resulting in the TEL/AML fusion gene. In adults we often find the BCR/ABL fusion gene, which is associated with poor prognosis. The MLL/AF4 fusion gene is also common in ALL.

1.2.4 Chronic Lymphocytic Leukemia (CLL)

CLL is a disease of the elderly with a high incidence in the Western world. In CLL, mature appearing B-lymphocytes accumulate in blood, bone marrow, liver, spleen and lymph nodes. The disease is biologically very heterogeneous. This heterogeneity is reflected in the diverse clinical courses the disease can take. In benign cases, patients may never receive treatment. In other cases, however, the disease can take a very aggressive course with a short survival time or the disease may be indolent for years and then transform into an aggressive form. Rai and Binet independently developed two clinical staging systems for better risk stratification in newly diagnosed CLL (Rai *et al.* 1975; Binet *et al.* 1981). These two staging systems are based on physical examination and standard laboratory test (i.e. lymph node status, platelet and red blood cell counts; enlargement of lymph nodes, spleen and liver). The discovery of biological markers in CLL like the *IGHV* mutational status and chromosomal abnormalities as determined by fluorescence *in situ* hybridization (FISH) has further improved the prediction of individual CLL cases.

1.3 Oncogenesis

Today there is a consensus that cancer is a multistep process of genetic alterations that give rise to the transformation of normal human cells into malignant cells. In 1914, Theodor Boveri, a German biologist, published his book *Zur Frage der Entstehung Maligner Tumore* (Gustav Fisher Verlag Boveri 1914) where he proposed the somatic mutation theory of tumorigenesis. Later, genetic alterations like chromosomal translocations, deletions or point mutations were experimentally proven to be associated with cancer. Therefore Boveri is generally acknowledged as the father of the somatic mutation theory of oncogenesis (Barrett 1993). Many other scientists have contributed and expanded this theory. In studies on patients with hereditary or sporadic retinoblastoma Knudson discovered that these patients need two hits in both alleles of a protective (tumor suppressor) target for the retinoblastoma to evolve (Knudson 1986). Knudson developed his two-hit hypothesis long before the retinoblastoma gene *RB1* was discovered and its function in cell division, proliferation and regulation of cell death was characterized (Chau and Wang 2003). This two-hit model is the basis for today's understanding of tumor suppressor genes. A tumor suppressor gene can still be protective in the heterozygous state (i.e. if only one copy of a tumor suppressor gene is non-funtcional). But if the second allele of the tumor suppressor is inactivated due to a mutation, often point mutations or deletions, its protective effect on the cell is lost. Quite often, the phenomenon of loss of heterozygosity at the tumor suppressor gene locus is a sign that the second hit has occurred.

In addition to tumor suppressors, a second type of genes, the so-called oncogenes, are involved in tumorigenesis. Oncogenes are mutated proto-oncogenes. Proto-oncogenes typically stimulate cell division and block cell differentiation in normal cells. Mutations like point mutations, deletions, gene amplifications or translocations in oncogenes are always dominant and lead to abnormal cell growth due to a gain of function. In 1976 Peter Novell proposed the clonal evolution model for tumor cell populations (Nowell 1976). He hypothesized that a normal cell which acquired a genetic alteration will have a growth advantage compared to its normal neighboring cells resulting in a clonal expansion of genomic instable cells, which acquire further alterations. Cells with most growth promoting additional alterations will be selected, eventually resulting in rapidly multiplying cancer cells (Nowell 1976). He considered leukemias to be early in this evolutionary process and solid tumors to be late. Michal Renan tried to answer the question how many alterations – that occur after birth - a specific tumor would exactly require to evolve (Renan 1993). Mathematically he showed that there is a group of tumors resulting from many somatic mutations like the prostate cancer (12 mutations), which usually evolve at older age, and a group of tumors that require fewer somatic mutations, like tumors affecting the brain and nervous system (4 mutations). These tumors are often found in younger people and often associated with an inherited predisposition (Renan 1993). Renan also hypothesized that hematopoietic malignancies require fewer genetic alterations than solid tumors, as malignancies of the hematopoietic system don't need to acquire properties that allow the malignant cells to metastasize. This hypothesis held true for leukemias and Hodgkin's disease but not for multiple myeloma (Renan 1993).

CML is an example for the progression of a relatively indolent disease to a more aggressive state due to additional alterations. At the start of CML, almost all patients have a t(9;22)(q34;q21) translocation resulting in the BCR/ABL fusion protein. With disease progression, they accumulate additional genetic alterations leading to a more aggressive form of the disease. This "multistep theory of cancer" (Vogelstein and Kinzler 1993) was further refined by Hanahan and Weinberg. Based on common molecular, biochemical and cellular traits, they proposed that six capabilities (hallmarks) are shared by most or all human tumors: self-sufficiency in growth signals, insensitivity to growth-inhibitory signals (antigrowth), evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis (Hanahan and Weinberg 2000). Survival, proliferation and dissemination of tumor cells is possible through two enabling characteristics: the development of genomic instability and tumor-promoting inflammation (Hanahan and Weinberg 2011). A single genetic lesion can affect more than one of these processes and also the same process can be disrupted by several lesions.

1.3.1 Genetics of Leukemias

Hematopoietic stem cells and their progeny are very well studied as well as diverse leukemias. Kelly and Gilliland proposed that myeloid leukemias are the result of at least two broad classes of mutations: (1) Class I mutations resulting in an increased proliferation of HSCs and (2) class II mutations which cause impaired differentiation and/or apoptosis (Kelly and Gilliland 2002). So far, a great number of chromosomal rearrangements have been identified and many of them have been studied in mouse models. Transcription factors play a major role in normal hematopoietic development and are frequently affected by translocations in AML. Very common are translocations that affect the core binding factors like the translocation (8;21)(q22;q22) resulting in the production of the AML/ETO or RUNX1/RUNX1T1 fusion protein or the translocation t(15;17)(q24;q21) and its fusion protein PML/RARα, which is found exclusively in promyelocytic leukemia. From mouse models we know that fusion proteins like PLM/RARα and AML/ETO, which belong to the Class II mutations (impairing hematopoietic differentiation), do not, on their own, lead to the development of leukemia (Pollock *et al.* 1999; de Guzman *et al.* 2002). Typical examples for Class I mutations, which confer proliferative advantage, are constitutively activated tyrosin kinases like BCR/ABL or mutations in *N-RAS*, *K-RAS* or *FLT3*. These alterations alone are also not able to cause leukemia. They lead only to leukocytosis with normal maturation and cell function (Kelly and Gilliland 2002). An AML phenotype (increased proliferation and impaired differentiation) could be observed when both types of mutations collaborate as it was shown for mutated *FLT3* and *PML/RARα* (Kelly *et al.* 2002).

However, the simplistic concept of class I and class II mutations is not sufficient to explain the plethora of mutations we find in AML. The functional grouping of mutations into two classes cannot always be upheld. From bone marrow transplantation models in mice we know that certain *bona fide* class I mutations as well as certain class II mutations on their own can induce aggressive leukemias. The same applies for the two-hit concept of a tumor suppressor as mentioned earlier, in CLL we find cases where the tumor suppressor gene *TP53* is only monoallelicly affected but these cases have an equally poor disease course as do cases with biallelic disruption of the *TP53* gene. We still do not understand how all the alterations in cancer interact, how cancer is initiated and which factors influence individual disease course.

With next-generation sequencing (NGS) technologies we are, for the first time, able to systematically identify genomic lesions in cancer in an unbiased fashion across the entire genome. In addition, this technology can provide us with information about the chronological evolution of alterations during disease course. The information generated by next-generation sequencing projects of cancer genomes has already transformed our view of cancer and will undoubtedly provide us with critical information in the future.

NGS experiments produce terabytes of data within a short time. The analysis of these data is very challenging. One of the biggest challenges in this analysis is to distinguish between driver mutations, which contribute to leukemogenesis, and passenger mutations, which just happened to be there in the transformed cell.

1.4 Biomarkers in Oncology

"Biomarker" is a very broad term. A biomarker can be any measurable indicator of a physiological or pathological cellular or organismal process. The working group of the National Institute of Health (NIH) defined biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Biomarkers Definition Working Group 2001; Oldenhuis *et al.* 2008). The World Health Organization (WHO) further expanded the term by including incidence and outcome of disease: "any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease" (WHO International Programme on Chemical Safety 2001). Therefore a biomarker can be everything from a blood pressure measurement, the concentration of metabolites or other substances in body fluids or the result of cytogenetic or molecular analyses. Most important for oncology are the detection and evaluation of prognostic and predictive markers. A prognostic biomarker is able to provide information about the overall cancer outcome regardless of therapy. However, based on presence or absence of such a prognostic marker a treatment can be selected (Oldenhuis *et al.* 2008). A predictive biomarker, in contrast, provides information about the expected effect of a given therapy. A marker that is predictive can at the same time be a therapeutic target, like the BCR/ABL fusion protein in CML. Next generation sequencing technologies enable us to detect many potential markers but they need to be validated thoroughly before they can be used in clinical applications: "a clinical useful prognostic marker must be a proven independent, significant factor, that is easy to determine and interpret and has therapeutic consequences" (NIH consensus conference 1991; Oldenhuis *et al.* 2008). The number of mutations that occur in a certain tumor does not make that task easier. Here again the discrimination between driver and passenger mutations is very important for the identification of prognostic and predictive factors in cancer.

1.4.1 Prognostic Biomarkers in CLL

The clinical staging systems in CLL developed by Binet and Rai are still of great relevance to predict outcome in newly diagnosed patients. However, predicting the individual disease course in CLLs still remains elusive because of the great heterogeneity. Certain chromosomal aberrations and the mutational status of immunoglobulin heavy variable chain region are still the most relevant and best studied prognostic biomarkers in CLL.

1.4.1.1 *IGHV Mutational Status*

For a long time the general opinion in the CLL community was that the disease derives from naïve B-cells that do not have mutations in the immunoglobulin heavy chain variable genes (*IGHV*). Nowadays we know that CLL originates from two groups of antigen-experienced mature B-Cells, which can be distinguished according to the mutational status in their *IGHV* genes. The initial suggestion of Schroeder and Dighiero that CLL may carry mutated *IGHV* genes has been confirmed by several groups (Schroeder and Dighiero 1994; Oscier *et al.* 1997; Fais *et al.* 1998). A key finding was the discovery that the mutational status of the *IGHV* genes

correlates with the patients' prognosis (Hamblin *et al.* 1999; Damle *et al.* 2002). Patients with mutated *IGHV* status (≈60 %) have significantly better prognosis and an indolent disease course as compared to patients with unmutated *IGHV* status who have a much more aggressive disease course. To answer the question whether both CLL types have different precursors, gene expression analysis revealed that -contrary to expectations - unmutated CLL (U-CLL) and mutated CLL (M-CLL) have a very similar profile suggesting a common precursor. The similarity of the gene expression profile of CLL and $CD27^+$ cells, a memory cell marker found in all B-CLL cases, implies that this common precursor is an antigen-experienced memory like B-cell (Klein *et al.* 2001; Damle *et al.* 2002; Klein and Dalla-Favera 2010). In normal B-Cell development, B cells enter lymphoid tissues where they are either activated by T cells (T cell dependent immune response) and pass through germinal centers where they experience somatic hypermutation in their V-genes, or they are activated in a T cell independent fashion and do not traverse the germinal center (no hypermutation). Both B cell types differentiate into plasma or memory cells. Therefore, U-CLLs might bypass germinal center somatic hypermutation through a T cell independent mode of activation (Chiorazzi *et al.* 2005; Klein and Dalla-Favera 2010). Lately, whole genome bisulfite sequencing in CLL revealed a DNA methylation signature that distinguishes three subtypes (Kulis *et al.* 2012). In contrast to gene expression profiling, methylome analysis using whole-genome bisulfite sequencing, revealed that U-CLLs have methylation pattern similar to pre-germinal center B cells and the methylome of M-CLL was similar to memory B cells. However, a third methylation pattern was recognized suggesting that this CLL subtype derives from an experienced, germinal center–independent B cell with low levels of somatic hypermutation (Kulis *et al.* 2012). This data demonstrate that we still understand only very little about CLL progenitor cells. It also shows out that there is more than one progenitor for CLL, as one can also deduct from the variable clinical course of CLL, suggesting that this disease should be divided in different subtypes. The *IGHV* mutational status had great impact on the understanding of the biological background of CLL and certainly it is a valuable prognostic tool.

To define the *IGHV* status of a patient sample, the *IGHV* status of the dominant clone needs to be evaluated by multiplex polymerase chain reaction (PCR). Then the dominant clone is sequenced and its sequence is compared to the germline sequence of the *IGHV* gene. A cut-off value of 98 % homology was defined to distinguish between unmutated (≥98 %) and mutated cases (<98 %) (Hamblin *et al.* 1999). As this technique is so laborious and expensive it cannot be performed in every laboratory. In addition, a mutational status cannot be precisely identified for each case. Therefore, this marker does not fulfill the NIH criterion of a prognostic marker which should be easy to assay (NIH consensus conference 1991; Oldenhuis *et al.* 2008). The expression of CD38 and ZAP-70 has been shown to be surrogate markers for *IGHV* mutational status, which can be detected by flow cytometry or real time PCR. For both surrogate markers there is no full concordance with the *IGHV* status and each of them has emerged as independent prognostic marker for CLL.

1.4.1.2 *Chromosomal Aberrations*

80 % of all CLL cases have chromosomal deletions and/or amplifications when assayed by FISH using a defined panel of probes. Patients' survival correlates with genomic aberrations from the most adverse aberrations to more favorable in the following order: deletion 17p13 > deletion 11q22-q23 > trisomy 12 > no aberration > deletion 13q14 as sole abnormality (Dohner *et al.* 2000). The deletion 17p13 affects the tumor suppressor *TP53,* which explains the dismal prognosis associated with this alteration. The deletion on chromosome 11 affects the *ATM* gene, which plays a crucial role in the DNA damage repair pathways. Multivariate analysis revealed that the FISH markers detecting 17p13 deletions and 11q22-q23 deletion are independent prognostic factors. They are independent of *IGHV* gene mutational status, age, leukocyte count and lactate dehydogenase level (Krober *et al.* 2002).

However, the predictive power of the FISH status or the *IGHV* mutational status is limited and these markers cannot fully explain the heterogeneity and the biology of CLL. Identifying new markers is important for prognosis prediction and for the identification of novel diagnostic targets.

1.5 Sequencing Technologies

A decade ago the human genome project was completed. The 3 billion base pairs of the human genome were sequenced using the conventional capillary-based Sanger method (see 3.5.2). In total, the costs for the whole human genome amounted to 3 billion US-Dollars. Simultaneous sequencing was limited by the number of capillaries (usually 96) of the automated sequencing machines. This automated Sanger-sequencing technology is also sometimes called “first-generation” sequencing technology.

1.5.1 Next Generation Sequencing Technologies

Since that time, alternative sequencing strategies have been developed to lower the sequencing costs and to dramatically increase the output of the sequencing machines. For this massive parallel sequencing of DNA, several new platforms were developed. This “second-generation” of sequencing technologies, which are commonly called “next-generation” sequencing technologies, have significantly changed genome research. NGS technologies are based on PCR clonally amplified templates and sequence data is achieved in a cyclic process.

With the new developed sequencing platforms, the costs for sequencing dropped dramatically. The National Genome Research Institute, a division of the National Institute of Health, has tracked the costs of DNA sequencing performed at the sequencing centers funded by the institute (http://www.genome.gov/sequencingcosts/). These data is freely available and demonstrate the remarkable improvements in DNA sequencing technologies. According to these data, we are close to the 1000 Dollar human genome. From the graph one can easily see the time point in 2008 when the sequencing centers switched from Sanger sequencing to NGS technology.

The first commercially available next-generation sequencing platform was launched in 2005 by 454 Life Sciences (Margulies *et al.* 2005), which was purchased by Roche in 2007. Shortly thereafter, Solexa’s Genome Analyzer and the SOLiD platform from Agencourt were released. These three platforms became the most commonly used sequencing platforms. In 2006, Applied Biosystems acquired Agencourt and Solexa was purchased by Illumina. All three platforms use adapter-ligated fragment libraries.

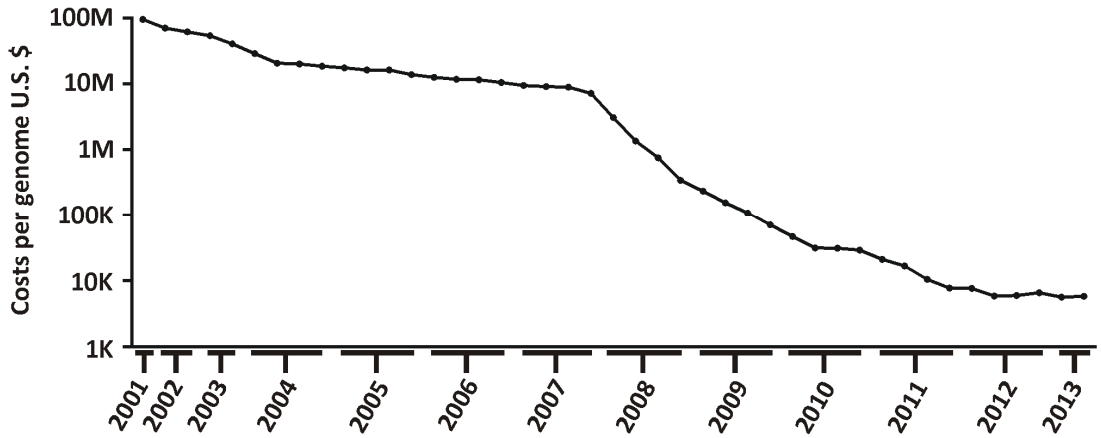


Figure 1 Costs per genome. Within twelve years the sequencing costs per genome dropped from 95 million to 5.826 U.S. $ (Logarithmic scale on Y axis). Data adapted from the National Genome Research Institute (https://www.genome.gov/sequencingcosts/).

1.5.1.1 *454 Pyrosequencing*

The 454 platform is based on an emulsion method for DNA amplification and an instrument using pyrosequencing technology (Margulies *et al.* 2005). Pyrosequencing relies on the detection of pyrophosphate, which is released during nucleotide incooperation by a polymerase. The complementary strand of a single stranded DNA is synthesized by adding only one of the four bases to the sequencing reaction at a time. The pyrophosphate, stoichiometrically released, when a base is successfully incorporated serves as fuel for a set of downstream reactions that generates visible light produced from a luciferase-catalyzed reaction: (1) one of the four dNTPs is added complementary to the DNA template using a sequencing primer (2) the released pyrophosphate (PPi) is converted into ATP by ATP sulfurylase in the presence of adenosine 5´ phosphosulfate (APS) (3) this ATP in turn provides the energy for the luciferase mediated conversion of luciferin to oxyluciferin. The light emission is detected by a CCD camera and is proportional to the number of incorporated deoxynucleotides of one type. Before starting the next sequencing reaction, unbound dNTPs needs to be removed. One can read the sequence when the type of dNTP added to each reaction is compared with light emission. It is important, not to use dATP as sequencing substrate as it is also a substrate for luciferin which would result in light emission, therefore dATPαS is used instead.

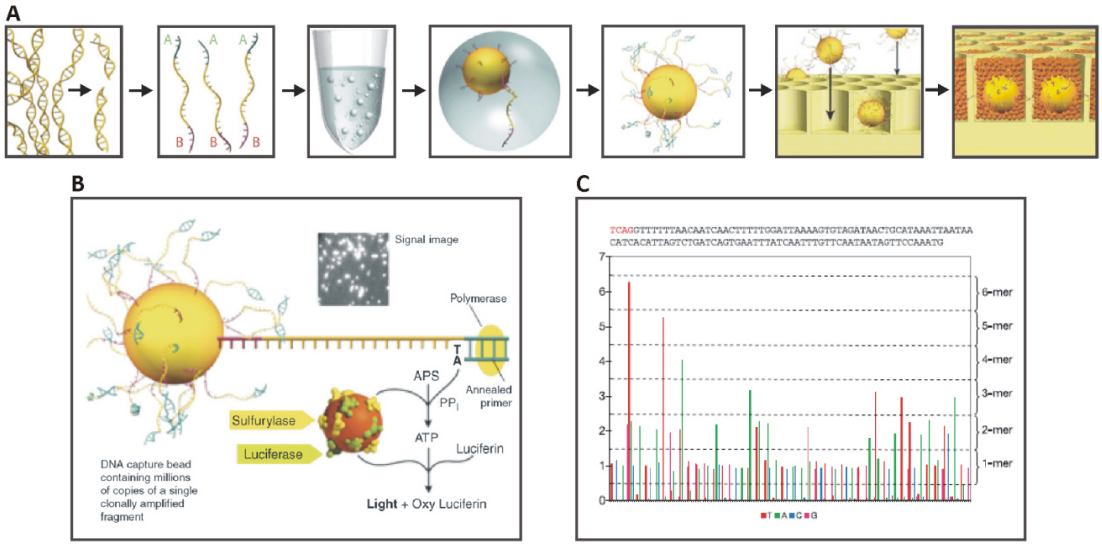


Figure 2 454 Sequencing Technology. A) Genomic DNA is fragmented before the sequencing adapters are ligated to each of the fragments. Thereafter, single stranded fragments are bound to beads which provide oligonucleotides on their surface that are complementary to the ligated adapters. Only one fragment is bound to a bead and each bead is captured in a droplet of PCR reagents in oil emulsion. The fragments are amplified so that each bead carries millions of copies of the initially bound fragment. The beads that carry single-stranded DNA copies after denaturation are spread across the wells of a nano titer plate in order to perform pyrosequencing. The reagents for sequencing are deposited into the wells in form of small beads that carry the reagents. B) The pyrosequencing process is summarized. Starting from the sequencing primer the first nucleotide is incorporated. Nucleotides are delivered in a sequential fashion. The released pyrophosphate is substrate of the sulfurylase enzyme which converts APS into ATP which itself provides energy for the conversion of luciferin into oxyluciferin catalyzed by luciferase. C) A camera detects light emission, which is proportional to the number of integrated nucleotides. Homopolymers have different signal values as indicated on the right. Images for this figure were taken from Margulies *et al* and Mardis *(Margulies et al. 2005; Mardis 2008a)*.

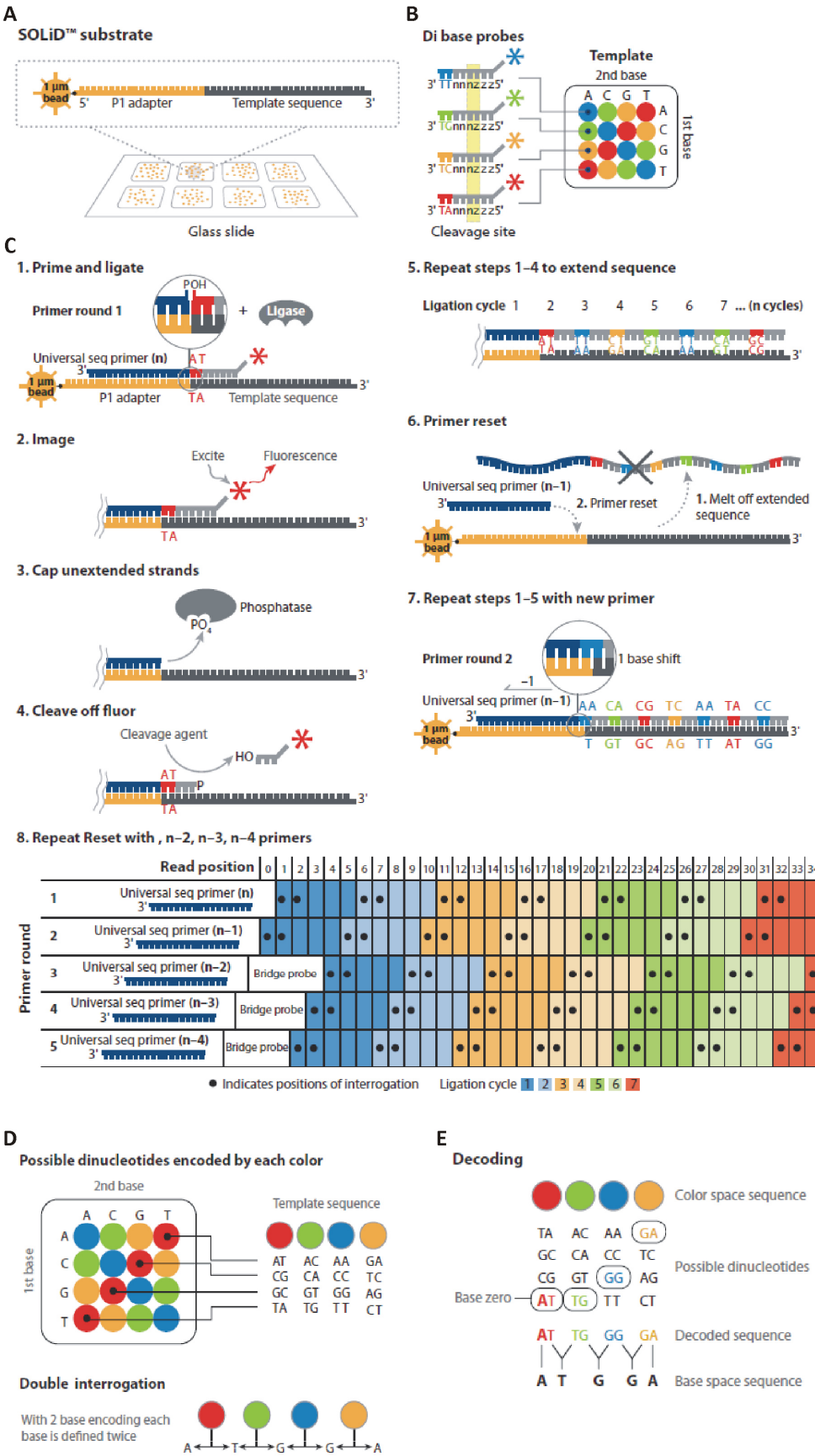
For the 454 sequencing, genomic DNA is sheared to generate a random library and adapters are added to the fragments. By limiting dilution, single DNA fragments are isolated and bound to beads (one fragment per bead) which have oligomers attached on their surface that are complementary to the adapter sequence of the DNA fragments. Within the droplets of an emulsion the fragments are clonally amplified. Each bead carries ten million copies of a unique DNA template (Margulies *et al.* 2005). For subsequent sequencing the emulsion is broken and the DNA strands are denaturated. The beads, which now carry single-stranded DNA clones, are deposited into wells of a fibre-optic slide. Enzymes (sulfurylase and luciferase) which are necessary for pyrosequencing are immobilized on smaller beads which are also deposited into each well. In the sequencing instrument the reagent solutions and the four nucleotides are flushed over the plate in a sequential fashion. If a nucleotide is attached by the polymerase, it will emit a burst of light. This light emission is directly proportional to the number of a particular nucleotide incorporated. After each cycle a washing step with apyrase, which degrades nucleotides, is performed to ensure that no nucleotides remain in the wells. Initially the GS-FLX instrument produced an average read length of ~250 bp per bead and a throughput of ~100 Mb of sequence data per 7-h run (Mardis 2008a). The long fragments generated make it easy to assemble fragments into large contigs. A drawback of this technique is the difficulty in accurately reading homopolymeric sequence stretches, because the linearity of signal

response is not well preserved for more than 8 identical nucleotides in a row (Margulies *et al.* 2005).

1.5.1.2 *SOLiD Sequencing*

Like in the 454 technology, in SOLiD (Sequencing by Oligo Ligation Detection) sequencing the templates are also amplified with an emulsion-based PCR approach. For sequencing, the fragments are attached to a class slide. Instead of polymerization, this technology uses sequential ligation of dye-labeled 8-mer oligonucleotides of which each probe assays two base positions at a time (Breu 2010). The ligation process starts with the annealing of a universal sequencing primer. The 8-mer oligos used for ligation are designed in a way that always two positions correspond to a specific fluorescent dye. As there are four different bases we have 16 possible dinucleotide combinations. Therefore each of the four fluorescent dyes encodes four different combinations of bases. At first an octamer will bind next to the sequencing primer where it becomes ligated. With a camera the fluorescence of the ligated oligo is recorded. In the next step the fluorescent label is chemically cleaved together with the last three bases of the octamer. By repeating these steps the sequence of a template is determined up to a read length of 35 bases (Mardis 2008a). At that point, only every fifth base is labeled. In the next step the extended primer is stripped from the template. To fill the gaps in the sequence, the whole ligation process is repeated by using a universal sequencing primer which is shifted one or more bases to the left. Several rounds are performed until the complete sequence is known. Alternatively the gaps can be filled up by using octamers where the positions of the labeled bases are shifted (Shendure and Ji 2008). In the example in Figure 3 the labeled bases are at position one and two. After five ligation rounds the data can be combined to generate a complete sequence. The ligation steps can be performed from both sides of a fragment so that one can also obtain paired-end information. The advantage of this technique is the high accuracy of base calling, as each base is queried twice. It is easier to discriminate SNPs from base calling errors as a SNP is characterized by a two color change. Initially it took 5 days to generate 3000 Mb of paired-end data with a read length of 35 bp (Mardis 2008a).

Figure 3 SOLiD Sequencing Technology. A) Fragments which were amplified by emulsion PCR are deposited onto a glass slide. B) Fluorescent-labeled 8-mers with dinucleotides encoded at position one and two. The cleavage site is between the 5^{th} and 6^{th} nucleotide. C) A universal sequencing primer is bound to the fragment prior to the first ligation step. After ligation the fluorescent signal of the ligated 8-mer is detected. After signal detection of the 8-mer by a camera, unextended strands are capped and the label is chemically cleaved. By repeating these steps, the sequence is extended. When a strand is completed the extended nucleotides are melted off. The entire process is repeated four more times with sequencing primers that are shifted one or more bases to the left. Every base of the template is interrogated twice as each fluorescent label encodes a two base combination. D) Color code for the 16 possible dinucleotides. E) As each base is read twice, the accuracy in base-calling is improved. Images for this figure were taken from Mardis 2008 and Breu 2010 (Mardis 2008b; Breu 2010).


A
SOLiD™ substrate
1 µm bead
5'
P1 adapter
Template sequence
3'
Glass slide
B
Di base probes
Template
2nd base
A C G T
1st base
3' TTnnnzzz5'
3' TCnnnzzz5'
3' TCnnnzzz5'
3' TAnnnzzz5'
Cleavage site
C
1. Prime and ligate
Primer round 1
POH
Ligase
Universal seq primer (n)
AT
TA
2. Image
Excite
Fluorescence
3. Cap unextended strands
Phosphatase
4. Cleave off fluor
Cleavage agent
HO
P
5. Repeat steps 1–4 to extend sequence
Ligation cycle 1 2 3 4 5 6 7 ... (n cycles)
6. Primer reset
Universal seq primer (n–1)
2. Primer reset
1. Melt off extended sequence
7. Repeat steps 1–5 with new primer
Primer round 2
1 base shift
–1
Universal seq primer (n–1)
AA CA CG TC AA TA CC
T GT GC AG TT AT GG
8. Repeat Reset with , n–2, n–3, n–4 primers
Read position
Primer round
Universal seq primer (n)
Universal seq primer (n–1)
Universal seq primer (n–2)
Universal seq primer (n–3)
Universal seq primer (n–4)
Bridge probe
Indicates positions of interrogation
Ligation cycle
D
Possible dinucleotides encoded by each color
2nd base
Template sequence
1st base
AT AC AA GA
CG CA CC TC
GC GT GG AG
TA TG TT CT
Double interrogation
With 2 base encoding each base is defined twice
E
Decoding
Color space sequence
Possible dinucleotides
Base zero
Decoded sequence
Base space sequence

1.5.1.3 *Illumina Genome Analyzer*

This is the sequencing technology used in the current work to analyze 25 CLL exomes and their matched normal controls. As the technique is already described in material and methods section (3.7) it will be only briefly reported here. This technology has its origins by the work of Pascal Mayer and colleagues (Adessi *et al.* 2000). Like the 454 and SOLiD platform, it is also based on adapter flanked libraries, which are initially clonally amplified. In contrast to the other methods described, the fragments are bridge amplified while they are attached onto a glass slide, the flow cell. The fragments remain immobilized on the flow cell while the reagents flow through. Elongation starts after binding of a universal sequencing primer. All four nucleotides, labeled with different dyes, and the polymerase are run through the flow cell. The nucleotides have a modified 3'-OH group which allows the incorporation of only one nucleotide at a time. An image is taken before the moiety at the 3'-OH as well as the fluorescent dye is removed from the incorporated nucleotide. As the DNA synthesis is terminated after adding one nucleotide but continued when the blocking group is removed, the nucleotides are called "reversible terminators" and the method referred to as "sequencing by synthesis". Initially this process could be repeated to generate 32 to 40 bp long reads (Mardis 2008a). In one run 1300 Mb could be generated in 4 days (Mardis 2008a). Incomplete cleavage of fluorescent dyes or 3'-OH terminators can cause signal decay or dephasing (loss of synchronicity) which is limiting for read length (Shendure and Ji 2008). With the improvement of the technology, the read length could be extended. In this work we performed paired-end sequencing with a read length up to 80 bp. With this technology a high output per run can be obtained. The most common error type of this technique is substitution (Shendure and Ji 2008).

1.5.2 Third Generation Sequencing

As there is always a next generation, the current generation (second generation) which we call now "next generation" is going to be replaced by "third-generation" sequencing technologies. Characteristic for second generation sequencing are the clonal amplification of single molecules and the sequence identification by alternating washing and scanning operations. These characteristics are also the major drawbacks of second generation sequencing technologies: (1) the clonal amplification based on polymerases, introduces PCR errors in the template and also amplification bias (2) all the clonally amplified fragments need to be synchronously prolonged, if this is not the case we see dephasing, especially at the end of a read; for this reason the read length is limited (3) the "wash-and-scan" technology takes time and requires large amount of sequencing reagents. By sequencing a single molecule without "wash-and-scan", third generation sequencing will overcome these drawbacks.

The Ion Torrent sequencer is one of two technologies on the edge of second and third generation sequencing technologies. Instead of using cameras this technology measures the pH change as a result of hydrogen ions during DNA synthesis. Similar to the 454 technology it also has problems with homopolymers and it still uses clonally amplified DNA molecules in the individual sequencing reactions. The first commercial available technology that can image

single DNA molecules is the Helicos Genetic Analyser Platform. Both Ion Torrent and Helicos machines still use a “wash and scan” technique. Third generation sequencing strategies overcome “wash and scan” and they are based on single molecule level, they can be subdivided into three categories: (1) observation of a DNA polymerase how it synthesizes DNA in real-time (sequencing by synthesis) (2) individual detection of bases as they pass through a nanopore (nucleotides cleaved of by endonuclease) (3) direct imaging of individual DNA molecules for example by transmission electron microscopy (Schadt *et al.* 2010). Most of these third generation sequencing technologies are still under development but might have the potential to replace the technology we currently use.

1.5.3 Data Analysis

The technical progress in NGS sequencing technologies is revolutionizing cancer research. At the same time these new technologies produce an overwhelming flood of data providing computational challenges: (1) Availability of computational power and processing speed to deal with the quantity of data. (2) Alignment and assembly of short reads to the reference genome. (3) Simultaneous analyses of matched tumor and non-tumor samples to identify cancer related variants.

With the development of NGS techniques also a variety of data analysis tools were developed. For each approach, the appropriate tools need to be selected and combined to generate an analysis pipeline to produce valid results. The hundreds of millions of sequence reads generated by a NGS platform need to pass several quality controls and they need to be mapped to a reference genome prior to variant calling. Variants can be detected with one of the numerous variant callers that have been developed so far.

1.6 Aim of Project

It was the aim of this work to identify novel genetic lesions in 25 well-characterized CLL relapse samples using next-generation sequencing technologies. For this purpose, an exome enrichment strategy was established in the laboratory. In addition, a bioinformatics analysis pipeline was set up to detect single nucleotide variants (SNVs), small InDels and to derive quality measurements like exome enrichment efficiency, coverage and sequence quality. Candidate genetic lesions identified in the NGS experiments were validated using Sanger sequencing.

2 Materials

2.1 Reagents

Reagent	Company
Agencourt® AMPure® XP beads	Beckmann Coulter, Brea, CA, USA
Bromphenol blue	Roth, Karlsruhe, Germany
ddH_2O	Millipore, Eschborn, Germany
dATP	NEB, Ipswich, MA, USA
	(NEB Next DNA Sample Prep Reagent Set1)
DNA-molecular weight marker VI	Roche Diagnostics, Mannheim, Germany
dNTP mix 10mM each	NEB, Ipswich, MA, USA
	(NEB Next DNA Sample Prep Reagent Set1)
Dynabeads MyOne Streptavidin T1	Invitrogen Dynal, Oslo, Norway
Ethanol	Merck, Darmstadt, Germany
Gel Red™	Biotium, Hayward, California, USA
Isopropanol	Merck, Darmstadt, Germany
Magnesium chloride ($MgCl_2$ 25 mM)	Roche Diagnostics, Mannheim, Germany
Nuclease free water	Ambion, USA
Oligonucleotides	Metabion, Munich, Germany
Orange G	Sigma-Aldrich, Steinheim, Germany
PE Adapter Oligo Mix	Illumina, San Diego, CA, USA
PE-PCR Primer 1.0	Illumina, San Diego, CA, USA
PE-PCR Primer 2.0	Illumina, San Diego, CA, USA
3100 POP6™-Polymer	Applied Biosystems, Foster City, CA, USA
Potassium chloride (KCl)	Merck, Darmstadt, Germany
Potassium dihydrogen phosphate (KH_2PO_4)	Merck, Darmstadt, Germany
Q-Solution	Qiagen, Hilden, Germany
RLT buffer	Qiagen, Hilden, Germany
Sodium acetate anhydrous (NaOAc)	Biomedicals, Illkirch, France
Sodium chloride (NaCl)	Merck, Darmstadt, Germany

Reagent	Company
Sodium hydrogen phosphate ($Na_2HPO4*2H_2O$)	Merck, Darmstadt, Germany
Sucrose	Sigma-Aldrich, Steinheim, Germany
Ultra Pure™ Agarose	Invitrogen, Carlsbad, CA, USA
Water for on-line analysis (sequencer)	Merck, Darmstadt, Germany

2.2 Enzymes

Enzymes	Company
F&P Super-Polymerase	Bio&Cell, Feucht, Germany
Herculase II Fusion Enzyme with dNTPs	Agilent Technologies, Santa Clara, CA,USA
Klenow Fragment	NEB, Ipswich, MA, USA (NEB Next DNA Sample Prep Reagent Set1)
Klenow Fragment (3'→5' exo$^-$)	NEB, Ipswich, MA, USA (NEB Next DNA Sample Prep Reagent Set1)
T4 DNA Ligase	NEB, Ipswich, MA, USA (NEB Next DNA Sample Prep Reagent Set1)
T4 DNA polymerase	NEB, Ipswich, MA, USA (NEB Next DNA Sample Prep Reagent Set1)
T4 Polynucleotide Kinase (PNK)	NEB, Ipswich, MA, USA (NEB Next DNA Sample Prep Reagent Set1)

2.3 Kits

Kits	Company
Agilent DNA 1000 Assay	Agilent Technologies, Santa Clara, CA, USA
Agilent High Sensitivity Assay	Agilent Technologies, Santa Clara, CA, USA
Big Dye® Terminator v1.1 Cycle Sequencing Kit	Applied Biosystems, Foster City, CA, USA
cBOT paired end cluster generation kit	Illumina, San Diego, CA, USA
CENTRI Sep8 well strips	Princeton Seperations, NJ, USA
QIAamp DNA-mini kit	Qiagen, Hilden, Germany

Sure Select Human All Exon 50 Mb (V3)	Agilent Technologies, Santa Clara, CA, USA
Taq PCR Master Mix kit	Qiagen, Hilden, Germany
Tru Seq SBS Kit V5-GA (36 cycles)	Illumina, San Diego, Ca, USA

The flow cell for next generation sequencing is included in the cBOT paired end cluster generation kit as well as all buffers and enzymes for cluster generation. The TruSeq SBS Kit V5-GA (36 cycles) includes all enzymes and buffers for sequencing on the Illumina Genome Analyzer IIx. All cBOT and Genome Analyzer runs were performed by Stefan Krebs at the Gene Center in Munich.

2.4 Buffers and Solutions

Name	Components	
AE buffer	10 mM	Tris-HCl pH 9.0
(Qiagen)	0.5 mM	EDTA
Bromphenol blue loading dye for marker	15 g	Ficoll
	0.1 mg	Bromphenol blue
	ad 100 ml	ddH_2O
DNA molecular weight marker VI	200 µl (50µg)	Marker
	800 µl	ddH_2O
	300 µl	Bromphenol blue loading dye
1x Ligase buffer	66 mM	Tris-HCl pH 7.6
(supplied with NEBNext DNA Sample Prep Reagent Set1 for adapter ligation)	10 mM	$MgCl_2$
	1.0 mM	DTT
	1.0 mM	ATP
	7,5 %	Polyethylene glycol
1x NEBuffer 2	50 mM	NaCl
(supplied with NEBNext DNA Sample Prep Reagent Set1 for Klenow Fragment 3'→5' exo^-)	10 mM	Tris-HCl pH 7.9
	10 mM	$MgCl_2$
	1.0 mM	DTT

Name	Components	
Orange G loading dye, 5x (for agarose gel electrophoresis)	10 g 50 mg *ad* 50 ml	Sucrose Orange G ddH_2O
1x PBS (Phosphate buffered saline)	136 mM 2.5 mM 10 mM 2.0 mM	NaCl KCL $Na_2HPO_4*2H_2O$ KH2PO4 Aqua bidest Autoclave
1x Phosphorylation reaction buffer (supplied with NEBNext DNA Sample Prep Reagent Set1 for generating blunt ends)	50 mM 10 mM 10 mM 1.0 mM	Tris-HCl pH 7.5 $MgCl_2$ DTT ATP
5x Sequencing buffer (Supplied with BigDye Terminator v1.1)	400 mM 10 mM	Tris-HCl pH9 $MgCl_2$
10x TBE (Tris-Borate-EDTA) buffer	1.0 M 0.9 M 10 mM EDTA Purchased from Invitrogen and diluted 1:10 with dH_2O for gel electrophoresis	Tris Boric Acid EDTA
1x TE (Tris-EDTA)buffer	1.0 mM 1.0 mM Purchased from Invitrogen	Tris pH 8.0 EDTA

2.5 Oligonucleotides

Oligonucleotides used for validation sequencing and screening of genomic DNA were designed either with the programs ExonPrimer or Primer 3 plus. ExonPrimer is a Perl script written by Tim M. Strom (Institute of Human Genetics, Helmholtz Center Munich). It is linked to the UCSC Genome Browser in the section "Sequence and Links to Tools and Databases" on each gene description page. The design algorithm for the oligonucleotides is based on Primer 3, an open source primer design tool (Rozen and Skaletsky 2000). All primers were designed with an annealing temperature between 60 and 63 °C. A second UCSC based tool called In-Silico PCR was used to verify that there was just one, unique amplicon for each pair of primers.

2.5.1 Oligonucleotides for Validation of Single Nucleotide Variants

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
ABCC10-324F	P1	CAGTGTTCTTACTTAGGGCATTGTC	
ABCC10-324R	P1	AAACAAGCAGCCAGCACCTC	303
ARHGEF17-320F	P1	GATGGGACAGTGAAGCCTCC	
ARHGEF17-320R	P1	AAGTCTGCGTGTGGGCTAGG	336
C1orf27-319F	P1	GTCTTTTGATCAAGGGTTTAGCC	
C1orf27-319R	P1	AGCAGAATAAAAGCAGATTGTTCAC	347
CHD2-14F	P1	TTTGACAATTTGCATGGCTC	
CHD2-14R	P1	ACATGAATGAACAGGCACCC	312
DMP1-18F	P1	CAATAATGAAATCCATCTGTGAGG	
DMP1-18R	P1	AACATCATTTCCTTCCATTCC	378
FAR2-15F	P1	CCCATTACAAATGCTTTCCA	
FAR2-15R	P1	TGGATTTTCAGATTAGGTGTGTTC	399
FOLR4-321F	P1	CCCTAGACACTTGCCGATCC	
FOLR4-321R	P1	TTAATATAATAACAGGGCATGGCAC	314
GRHPR-13F	P1	ATCTGGTTGTCCCTAGCCTG	
GRHPR-13R	P1	CCACTGCTGAGAAGCAGACAG	416
IL28A-21F	P1	CTCACACCTGCTCTCCCTTC	
IL28A-21R	P1	GACGCTGCTCAGAGCTCAC	355
PCDHGA4-322F	P1	ACCACACCCGGCTGCTC	
PCDHGA4-322R	P1	TACCCGGGGAAGAAGATTCC	1065
RUFY1-323F	P1	GCTTGTTGCTCTGGTTGCTG	
RUFY1-323R	P1	CTCCTAAATGGCCCACCCTC	601
ACOX2_29aR	P2	TCCTTGAGGGAGCATTATGG	
ACOX2_29F	P2	AGCAATGCACAGGTCTTCCT	196
ADAMTS12_30F	P2	CCCTGCTTTCTCAGTGGTTG	
ADAMTS12_30R	P2	GAAGGCAGGGGCTGATG	348
CDH12_25F	P2	CCTGGATTCCACAATTCTACG	
CDH12_25R	P2	AAATACACAATGCATATTCCCC	360
CREBBP_325aF	P2	CTGTTTTCGCGAGCAGGT	
CREBBP_325aR	P2	TGCTCCGAGCTCCCG	220
METRNL_327F	P2	TCCTCACTGACTCTCTGTGGGTG	
METRNL_327R	P2	ATCACCAGCCCCGTGATG	522
OR4A47_369F	P2	GGTGATGGCCTATGACTGCT	
OR4A47_369R	P2	CTTTTTGCCTCCCTTTCTGA	359
PAPD5_22F	P2	TGAGTGTCAGTTAAAACAAGTTTCC	
PAPD5_22R	P2	TGCAACTGTGCTCACCAAAC	274

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
RB1_24F	P2	AGACAAGTGGGAGGCAGTGT	
RB1_24R	P2	GCAAGAAAAGATTATGGATAACTACA	341
SRD5A1_326F	P2	CCAGGTAAGTATTCACTAGCATCTCTG	
SRD5A1_326R	P2	CACAGAATAAACTGGAATTCAGCG	351
TCHHL1 _23F	P2	AAGCATCTGAACACAATGATCC	
TCHHL1 _23F	P2	TTTGTTTTAGCTGGTGTCTGG	532
ZNF227_370F	P2	CAGTCAGGCCATAGATTTTCG	
ZNF227_370R	P2	GATGGCGAAGATTCAAGCTC	278
ATXN2L_35F	P3	GTGCTCCCTAACTCTGGCTC	
ATXN2L_35R	P3	CTCCAAGACCCCTCACTCC	421
C2orf67_82F	P3	GCTCTACATGGAAAGTCCCAGA	
C2orf67_82R	P3	TGGGCTCCCCTAATTTCTTT	233
EBF2_331F	P3	CACATGTGGCCTGACTGTGC	
EBF2_331R	P3	AACTTTCTCCCAAAAGGCCC	325
MAGI2_34F	P3	TCATTCTCTCACCACCAGCC	
MAGI2_34R	P3	CACCCTTTCATTGCCCTG	441
MED12_38F	P3	TTCCTTCTTTTCTCCTGCCC	
MED12_38R	P3	TCAGCCACTTAGGTTGTCCC	291
RPS6KA3_37F	P3	GATAATTTTGCTATTCCTTTCACG	
RPS6KA3_37R	P3	GCAGAGATGTGAAGCACAGG	391
SCUBE2_33F	P3	GAGGCCAGATCACACATGG	
SCUBE2_33R	P3	TGGCCACTGCAAGAGCTAAG	253
XPO1_42aF	P3, 4, 14, 15	TGAACAGAAAAGAGGCAAAGAT	
XPO1_42aR	P3, 4, 14, 15	TTCATTTATTTTGTCCTGGACTC	249
ADAMTS20_40aF	P4	GCCAAGAAACGGAGGAAATTA	
ADAMTS20_40aR	P4	AAAATTTACATCAAAGAACCAGCA	218
CCDC111_49F	P4	GAGTCCCCTGATCCATTCTG	
CCDC111_49R	P4	TTCTGTACCCTTACATTTTGATAGG	734
CSF2RB_44F	P4	GTGAAGTCAGGGTTTGAGGG	
CSF2RB_44R	P4	ATGACTGAGGAAGGTCAGGC	492
CYB5D1_50F	P4	AGTACGTGCTGAGGAGCAAAG	
CYB5D1_50R	P4	ATCCCCGCTCTCCATTTC	458
CYFIP2_41F	P4	AACATCTGCCAGGACTCCAC	
CYFIP2_41R	P4	GCAAACTTTGTTTATGTCAAACC	280
EMR2_43F	P4	TCCTGTCCTGTTGTGTTCCA	
EMR2_43R	P4	CAGTGGTGATGGAGGATGTG	210
GRM6_45F	P4	ATGTGGTGAGGACTGTGTGG	
GRM6_45R	P4	CGAGGCAAGAGGAAGAAAGG	329
PPP1R1B_333F	P4	CAGGACAGCCGGATGGATAC	
PPP1R1B_333R	P4	GGATAGAGTGGGTTTCTGGGG	387
RBPMS2_47F	P4	AAGGAAGCGCTGTTGTCATC	
RBPMS2_47R	P4	CGAGATTCACTCCCGCTG	216
SF3B1_48F	P4	GTTGATATATTGAGAGAATCTGGATG	
SF3B1_48R	P4	TTTAAAATTCTGTTAGAACCATGAAAC	536
ARHGEF5_385F	P5	CCCTGCCTCTCCTAACCTCT	
ARHGEF5_385R	P5	AGCTGCAGGAAGTGAACACA	227
BRDT_79F	P5	TGGCAGTTTTTAAATGTTCCTG	
BRDT_79R	P5	GTGCTCCCATCAACAGAGGT	249
CAPZA3_56F	P5	GGAAATTGCAACATGCTGAG	
CAPZA3_56R	P5	AGGAAAATAACCATTCCCAAG	560
IFT52_57F	P5	TTGATTTAATTCTCTAGTCCTGACC	

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
IFT52_57R	P5	TGGTACTGAGAAGCTGGACTC	209
KAT5_55F	P5	AAGCTCCTGGTTGACCCTG	
KAT5_55R	P5	CTGTGAGGCACCGATAGTGG	450
MYH2_386F	P5	TGCAGTATTAACAGTAAGCGTGTGG	
MYH2_386R	P5	GGAAAATTCCTGGGTAACAAGTTAAG	279
MYH7_76F	P5	CCTGTCCTGTGCTCTTCCTC	
MYH7_76R	P5	CCCTCACTGCCAATCCTC	250
PCLO_53bF	P5	GCAGCAGAAAGGAAGAGAGC	
PCLO_53bR	P5	TCTTCATAGGCAGCATCAGC	249
TIAM2_77F	P5	AGAGACTGCAACGGACACCT	
TIAM2_77R	P5	AGTCCTCTGCAGGCACTTTG	181
TRIML1_387F	P5	CAATGGAAATCTCACATACCTAAGAGG	
TRIML1_387R	P5	GTGACGAGACCCAGAGGCTG	398
GRK7_86F	P6	GGGAGACCTCAAGTTCCACA	
GRK7_86R	P6	CTTGCCACCCTTCATCTCC	217
KIAA1843_84F	P6	TCTGTGAGTCCATCCTGCAC	
KIAA1843_84R	P6	AAAGGATGCCATGCTAATTTTT	249
KLRC4_85F	P6	CCCCTCTCTCAGTGCCTCTT	
KLRC4_85R	P6	GAAATGTTTCAAGGCGCTTC	202
ANKS1B_60F	P7	GTTGCTTCAAAAGCGACTCC	
ANKS1B_60R	P7	ACACATTCCATTCCCAGACA	485
BAZ1A_66F	P7	AACAGCCACATGGAGTAATTG	
BAZ1A_66R	P7	TGACCAAAATTCTGGACCAAG	292
BMP2K_71F	P7	CCAATGTGAAAGGAAGGAAC	
BMP2K_71R	P7	TAATGACCGACCATTCAACG	353
C12orf51_335F	P7	AGGGAAACAGTGATGCCTGC	
C12orf51_335R	P7	CATTCTGATTTTCATCCAGCTCC	264
COL18A1_336F	P7	AAGCATGTCCCACCCTCCTC	
COL18A1_336R	P7	TCTCAGGGACACTCTCCTGC	238
FRAS1_61F	P7	TGATAATTTAGCCTCCAGTCTCC	
FRAS1_61R	P7	CCTCAAAACAAAGAGTGCACAG	397
KIAA2022_63F	P7	TGTTCAATGATGAGGATTCTGTC	
KIAA2022_63R	P7	TGGAGGCCATCATGAATCTC	594
KLHL6_64F	P7	CTCAGCTCCTGCAATGGG	
KLHL6_64R	P7	TCAAGATTGGGCTCTCACAC	424
KLHL6_74F	P7	AGTGGTTAAGAGGGAAGAAACC	
KLHL6_74R	P7	GGACTGGAGGAGGGTGAGAG	441
LIMD1_73F	P7	GACCCTTCGCCAGCATC	
LIMD1_73R	P7	AGGAAGGATCCTCACAGGG	650
LRP1B_69F	P7	GAGTGGGAGTATCATTTGAGCC	
LRP1B_69R	P7	AAGCAAAGAGGGAAAGGAAAG	535
MARCH6_67F	P7	TTTCTGGGGCCCACTTTAG	
MARCH6_67R	P7	CAGCCTGCTTGAGAACTGG	338
NBPF4_72F	P7	CTGTGTGGCGTGGGTCAC	
NBPF4_72R	P7	GGTTGGAGCATCATGGATTT	489
NOX1_337F	P7	CGCAACCACTGTCTGGAGC	
NOX1_337R	P7	CCTCCACCTCCAACCTCAAAC	577
OTUD7B_70F	P7	AAGGGTGGCAAGGAGGAG	
OTUD7B_70R	P7	AGGTGGCATATGGTGGTAGG	511
ROBO1_62F	P7	GCCTTCAGGTAGAGTGAAGGAG	

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
ROBO1_62R	P7	TGTCAGTGCAATTATAACATCTAAGG	324
UNC13C_68F	P7	TTACAGTGATTCTCAGCTCTCTTTAC	
UNC13C_68R	P7	TGTGGTACTTGACTCATTTGCAG	616
ANKFN1_339F	P8	GACGCAGGGCAAGGAAGAG	
ANKFN1_339R	P8	CTAAAATTACCGGCGCCACC	237
CELSR3_252F	P8	CAGAACCCCGTCATGAACTC	
CELSR3_252R	P8	GGACCAAGGGTTTCCAGAAT	193
DIO2_338F	P8	AGAGGGTGAAGGGGAACCAG	
DIO2_338R	P8	CTGGTCCCCAGCATATGAGC	344
DOCK10_244F	P8	CATATTCTGATATGGAATACTGACTCG	
DOCK10_244R	P8	CCTCCCTCATGTGGTCATCT	178
FREM2_250F	P8	CGTCATCCAAGATGGTCACA	
FREM2_250R	P8	GGGAAAATGGCAATGACTGA	162
GRIA2_246F	P8	GAAAGATGAGATTTGTTCATATTGTG	
GRIA2_246R	P8	TTGGACTTCCGCACTCTAGC	238
LRFN5_247F	P8	CCTTCCTGGACTTCAAGCAC	
LRFN5_247R	P8	GGATCTCTCAGCAGGCAAAG	240
LRRC7_241F	P8	TGGTGACAAGCCATCAGATAA	
LRRC7_241R	P8	TTCATGCTTCACTACATGTGCTT	224
MYO6_249F	P8	AACCTCTTTGATAGACAAATGGTATT	
MYO6_249R	P8	CTTCAGAAGCACCAGCACAC	233
PASD1_243F	P8	CCACGTGAGTGTCTCATCAGTT	
PASD1_243R	P8	GGAAACTGAGGAATGCTGGA	170
PDE1C_340F	P8	GCTTTGAACAGAGAGCTCCAAC	
PDE1C_340R	P8	TTGGAAACGTGTTGCTTTGG	308
RBM46_253F	P8	ACTATGGGGCCACACCATT	
RBM46_253R	P8	TCAACTGCACCAGGCTTAAA	165
TNFAIP3_242F	P8	GAGAGCACAATGGCTGAACA	
TNFAIP3_242R	P8	GGATGATCTCCCGAAACTGA	202
TNS3_245F	P8	TCCGGTTAATTGTCTCCACA	
TNS3_245R	P8	CAGCAAGGACCACCCAGTAT	152
TTN_240F	P8	GGCAGCCCTATCATTGGTTA	
TTN_240R	P8	CACATCACTGGGGTCACTGT	186
ADAM8_97F	P9	CCCAAGAAGGACATGTGTGA	
ADAM8_97R	P9	CTTGCAGCCTGGTAGGATGT	294
ADAMTS14_99F	P9	TGGGTTGGACCTCACTCTCT	
ADAMTS14_99R	P9	AGGACACATGGACACTCACG	247
AMHR2_341F	P9	GCCTCTGCATTCACTCCCAC	
AMHR2_341R	P9	CACACCCCAGGATGTGTCTG	301
ARSK_93F	P9	CACCCTGGATTCTGAGTGAA	
ARSK_93R	P9	TCCTAGCATGAGAAGCAGCA	247
ASH1L_92F	P9	TTGGAGGATTCACTCCCAAC	
ASH1L_92R	P9	CCCTCTTACCTGCCAAAGTC	231
DNAH5_344bF	P9	GGGTGAGACGATTGAATTGG	
DNAH5_344aR	P9	TTTCCTGTATTTTAATGCTGAATCT	177
MRO_91F	P9	CTCGGCCTTTGTTTTGTTTG	
MRO_91R	P9	TAGGGTCTGGCGTCTCAACT	219
MYH2_371F	P9	CTGTGTCACTGTCAACCCCTA	
MYH2_371R	P9	GGGACGATCTCAAGGAATGTG	240
PTPRT_96F	P9	CAGAGGGGGAAAGTGTTCAA	
PTPRT_96R	P9	TCCACTTCAAGGAAAATCAGG	222

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
SCD5_98F	P9	AAGCTCAGGGACCATTGTTG	
SCD5_98R	P9	TCTTTCCCCTCCTCCCATAC	246
SCRN2_343F	P9	CCATCCAGGTGGGAAGAATG	
SCRN2_343R	P9	TCCTTCATATTTCCAGAGTCCCAC	350
SIPA1L1_95F	P9	CCTTGGTCTCTGGACCTCTG	
SIPA1L1_95R	P9	GGAGGCACTTTCACATCCAC	230
SKIV2L2_94F	P9	CTTCTGGGCTACGAGAGACA	
SKIV2L2_94R	P9	GGTGGGACTTAAAATCTCAAAGA	398
TOP2A_90F	P9	TGGCTCGATTGTTATTTCCA	
TOP2A_90R	P9	TGGCAAAGGTTCTTCTCCAT	224
AKR1C4_104F	P10	TTTCGTTGCTCCTTCAGGTT	
AKR1C4_104R	P10	TCATCTCCACACAATCCCATT	244
COL5A2_106F	P10	CACCTACCATTCTTTGGGAAA	
COL5A2_106R	P10	AAATGTTCGTGTCAAGATACCC	376
ERCC8_103F	P10	TGGGTGAGGGGTACAGTCAT	
ERCC8_103R	P10	TGCGTTTATTATGTGGCTTCA	249
KIF9_382F	P10	TCTTGCCCCAGTGGATTCAG	
KIF9_382F	P10	TGCAAAGGTCTGAGGATGGG	217
KLRC2_373F	P10	AAGCCTGTAAGGATGCGTAA	
KLRC2_373R	P10	CAAATGTATTATTCACTGAAGGAAGC	333
PRKCA_105F	P10	CCAGTTCCAAAGCAAACCAT	
PRKCA_105R	P10	TGGGGCGATATAATCTGGAG	201
RANBP2_102F	P10	GGAAAGCTGGGCTTTAGGAT	
RANBP2_102R	P10	CAAATGTTCTGCTTTCTTCAGTG	300
ARMC4_109F	P11	GGCAGCCTTTTTAGCATCAT	
ARMC4_109R	P11	GGCATTAAATGACAAGAAGGTGT	299
ATM_115F	P11	AAGGTCCTGTTGTCAGTTTTTCA	
ATM_115R	P11	GGCTGAGATTTTTGGGGTCT	241
EGR2_346F	P11	CCTCGCAAGTACCCCAACAG	
EGR2_346R	P11	CAGCTCCAGTGGACAAAGGG	529
FAM179A_113F	P11	TGGGGACAGGTATTTTTGGA	
FAM179A_113R	P11	CTACCTGGTGTCCTGGTTGG	295
GPR61_345F	P11	CCAGGGTCGCTGGACTAGG	
GPR61_345R	P11	ACACACCCACCAGCACAGAG	570
KCNJ5_112F	P11	ACTGGATGGGTGGACGATTA	
KCNJ5_112R	P11	TTCATTTCTGCCAGCTCCTT	746
MRE11A_110F	P11	GCATAAACACTGTGAATACTGAAGG	
MRE11A_110R	P11	TTCCCACTGTCAATTTGTTTAAGA	356
NRBP1_111F	P11	CCTGGCCTGACATCAGTGT	
NRBP1_111R	P11	AAATCCTCCATTTTAACACTCCA	285
SLC2A2_347F	P11	TCAGGGAGGGGCTTTCATTC	
SLC2A2_347R	P11	TGGAGGAAGTACAGTAGGGGATG	316
ZFP161_108F	P11	AACGCCTGGAAGGAGAATTT	
ZFP161_108R	P11	TGTCAGCAGCATCTCCAATG	387
ABCA9_121F	P12	GACGGCTTTGTTTCTCTGCT	
ABCA9_121R	P12	CCACCCCTAGACACTGTTTCA	296
ACSM2_407F	P12	AGTGAGGCCCCTAGGATTGT	
ACSM2_407R	P12	CGTTCCCATGTATCCTGGTT	187
ADAM17_125F	P12	TTGAAAGTAAGGCCCAGGAG	
ADAM17_125R	P12	CCAAAATCCCTGTGGAGAGA	188
CASP6_131F	P12	CCTTTGGATGTAGTAGATAATCAGACA	

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
CASP6_131R	P12	TCTGTGAGCTTCAGGATGGA	250
CDH12_124F	P12	CAAGGCCACAGATGCAGAT	
CDH12_124R	P12	CGCACTGAAGCTTTATGGCTA	214
CTAGE5_348F	P12	AACATGGCATTTTGGAAACTAGAG	
CTAGE5_348R	P12	GGCAACACAGCCATGCTAATTC	301
FLNB_130F	P12	CATACCTGGGTCTCCCTTCA	
FLNB_130R	P12	GGGAAGTGTGGCACGAGT	300
KIAA0240_118F	P12	CCAGGGCTCAGTAGTTGGTC	
KIAA0240_118R	P12	CTTCCAGAGGCAGCAAATGT	399
KIAA1586_122F	P12	TGATTCAATTTGTAATTTAGTGCAT	
KIAA1586_122R	P12	TCTTTGAAACTGTAGATGCCTCAT	296
KIAA1712_123F	P12	TTCCATCACAACAAAGAAAGAAAA	
KIAA1712_123R	P12	TCATCAGGTGCCAATCAGAA	213
LGR5_120F	P12	TGGATGCTAACCACATCAGC	
LGR5_120R	P12	GAAACTGTTAAAAGAGAGCAAAATCA	248
PACS2_132F	P12	AGAGAAAGCCCTTCAGAAACC	
PACS2_132R	P12	CAGGACTCAGGCCTCTTGAC	500
PAM_126F	P12	AGGGTGCAGAACATGAGAGG	
PAM_126R	P12	AAGTCCTTTTCATTGATGAATTTTT	273
PRPS1L1_374F	P12	CCAAGACGCCGAATATCAAA	
PRPS1L1_374R	P12	GCAACAAGCTTGGCAGAGAT	343
PTPRU_129F	P12	CCTGAGCGAGAATGATACCC	
PTPRU_129R	P12	CCAGCCCACCTGATATTCAT	214
RIMS2_128F	P12	GCTATCATATGAACCAAATGTAGGTA	
RIMS2_128R	P12	CAAAAGAAGAAAATTAACACAAAGG	288
RNASE2_119F	P12	CAATGCAATGCAGGTCATTAAC	
RNASE2_119R	P12	GGATACTGTGGAGGGTCTCG	300
BHLHB9_140aF	P13	GATTGCAATGGGTGTCCATA	
BHLHB9_140aR	P13	CTGTCCCGGTGAGTTCAAAG	232
C20orf54_146F	P13	GCTAGGTGGTGAAGCTGGAA	
C20orf54_146R	P13	ATGACCACCGTGAGGTAGGA	242
FAM32A_375F	P13	CTCTCCAGCAAATGGAAAGG	
FAM32A_375R	P13	AGCAGAAGATGCAAGGGTGT	562
GANC_143F	P13	TCACTTGGCCTTCTTTTGCT	
GANC_143R	P13	CCACGTGTGCATGGTAGAAC	250
GTSF1_144F	P13	TGTGCCTGGAATGCTGTAGT	
GTSF1_144R	P13	GGAGGGTGTTAGAGGGCAAA	204
HPSE_145F	P13	TTGTGGTGCCAATCTAACCA	
HPSE_145R	P13	CCATTGCCTAGTTCCCAAGA	220
IGFN1_147F	P13	CTCTGCCGTCTCTCCTGAAG	
IGFN1_147R	P13	CCACCCTGAAGTGGTATTCG	248
MYL12A_350F	P13	CTTTCCAGACTCTTTAATAGGCCC	
MYL12A_350R	P13	AAACTACTACTTAGGGCAGAAGCAGC	524
OR5W2_139F	P13	TTGATCGGTACAAGGCCATC	
OR5W2_139R	P13	AATTGCTAGATTGTGGCTGAG	720
ADAMTS3_159F	P14	CCCACTCCAGATCTCTGTGTC	
ADAMTS3_159R	P14	TTCACAGGTGTTTTCCTTTGG	371
ARSD_160F	P14	CGACTCCGAGCCCCTGTA	
ARSD_160R	P14	GCATGAACAGAACGGGAAAT	164
CCDC56_162F	P14	GCGGACTACAACTCCCAGAG	
CCDC56_162R	P14	GTGTCAGCTTCTCCCGAGTC	188

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
CCDC85A_157F	P14	GAGCACTCCAAGCACAGGA	
CCDC85A_157R	P14	GTGCTTGGGCAGCGTTTC	186
COL4A3_164F	P14	GGAAAGTTGCTGATGTGGAGA	
COL4A3_164R	P14	AAATTTACCCTTAGGTCCAGGAG	158
FLNC_155F	P14	CATCGTGAACACCCTGAATG	
FLNC_155R	P14	CACCGTACTTGATGGCAATG	158
FUBP1_156F	P14	GCCGCCATTTTCTTTCTTTC	
FUBP1_156R	P14	GCCAATTACCGTGAGCTTTC	247
GRIK1_376F	P14	ACAGAAACCGAACCCTGATG	
GRIK21_376R	P14	CGACCCATTTGCACAAAGAT	204
OR2L8_151aF	P14	GCAGCACCCACCTCACTGTA	
OR2L8_151aR	P14	CCTGAGACCTTAGGCAGAAAGTG	247
OR5J2_153F	P14	TGTGTTTCGGAGTGTTCATCA	
OR5J2_153R	P14	TCACAGAAGAAGTGGCTGACA	235
PLS1_150F	P14	TTTTGTGTTATTGGTAAAATGATGAAA	
PLS1_150R	P14	CAATGAATCAGTGACTAAAAGATCAAA	345
SLC6A11_154F	P14	AGCAGAGAGAGGGACCTTCC	
SLC6A11_154R	P14	AAAGATGGCCACGAAGAGAA	171
SLC6A5_152F	P14	GGAGATGCATGGACTCCTGT	
SLC6A5_152R	P14	TGCAGCTCTCTTAGGTTTCCT	363
SMCHD1_351aF	P14	TTGCAGGCGTGATCATTTTA	
SMCHD1_351aR	P14	AAAGCACAAGCTCGTGAACA	388
ZNF512B_158F	P14	TCCTGAAGGCTGACAGGTCT	
ZNF512B_158R	P14	TTCCACCTGCCAGATTTTTC	193
DNAH7_171F	P15	AAATTGCCATGGATGGTCTT	
DNAH7_171R	P15	CGATAACTGAGAATTTGTCCATCTT	166
FLJ32682_180F	P15	CCTTCTTGACTGTCCCGTTG	
FLJ32682_180R	P15	GGCAATACTCGAGCCACAAT	238
ITIH4_353F	P15	GCTTGCTGTGGGATCTGGG	
ITIH4_353R	P15	GTGTCATCCACAGGCAGCAG	694
KRT20_174F	P15	GGCAACAGAGCAAGATTCTG	
KRT20_174R	P15	GAACCGTGCTTCCTTTATCAAC	285
KRT4_178F	P15	CTAACCCAAGGAAAGCTGGA	
KRT4_178R	P15	GTAGGTGGCGATCTCGATGT	287
L2HGDH_175F	P15	CCCTGTTTCTCTTGTTCTGTAGC	
L2HGDH_175R	P15	ATGCCATACCTTGCCACACT	183
MYEF2_176F	P15	TGCATGTGAAAATGGTGAGTT	
MYEF2_176R	P15	GGCACAAAAACAAAGTGTTGC	280
PCLO_354F	P15	CAGCACATGCAGTGACATTGG	
PCLO_354R	P15	AAAGAAGCCACATTTTCATCCTG	605
RBM25_177F	P15	TGTGCTATGCACTTTAGGAGC	
RBM25_177R	P15	TCTAAACATGTTCATTGGTGAATC	511
SF3B1_352F	P15	TCTGCTGACAGGCTATGGTTC	
SF3B1_352R	P15	TGAAGAGAATACTCATTGCTGATTACG	725
SLIT2_172F	P15	AAAGGTTATCCTCTTCTCCTTCC	
SLIT2_172R	P15	AAAACTGAGGCCAGTAATGATTC	292
XPR1_173F	P15	CCAATTAGTGTCATCCCCACA	
XPR1_173R	P15	TGTACAACCTAGCAGCCTTGG	212
AKR1B1_355F	P16	CCAGCAGGTCTGTGAAGGAC	
AKR1B1_355R	P16	AATGGCAGGCAGATTGCTTC	217
ARHGAP18_187F	P16	CAGTACCATTTTCTCTGGTTTTG	

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
ARGHAP18_187R	P16	TGATCCACTTGAGACCTTACATTAAC	263
ASMTL_191F	P16	GGAGACGCTCCTGGATGA	
ASMTL_191R	P16	CAGTCACGACTACACGCTCCT	311
DEFB114_188F	P16	CATTGTCGTGTAGCTCTACAAATATC	
DEFB114_188R	P16	CCACACCTCTCTGCACTGG	519
DOCK1_189F	P16	CCCTCAAGTCTGGATGATAGAA	
DOCK1_189R	P16	CAACAACTCACACGACCTCAA	340
DST_186F	P16	CTGTGAAGCACTGTGGCTCT	
DST_186R	P16	CTGCCAGTTGCTCTTTCACA	315
LAMA5_193F	P16	AGCAGGCAGAGGTGCACTA	
LAMA5_193R	P16	CAGGGACAACACCACTCACA	250
OBSCN_190F	P16	GCTACCCTGCTCCCAGATTA	
OBSCN_190R	P16	TGTGCCACTGTGTACCAAGG	336
DNAJC13_194aF	P17	GCACATGGGACTGCTTATGA	
DNAJC13_194aR	P17	TCATCCACGTGGCACTTCTA	238
GRIN2A_357aF	P17	TCCGACGTCTACCTTCTTCC	
GRIN2A_357aR	P17	GAGAACAGCCTCGTCTTTGG	298
ICAM5_356F	P17	GATAGTGCATGTCAAGTGCTTAGG	
ICAM5_356R	P17	CCTTCGCATTCCTCTGGTCC	418
LAMA1_358F	P17	GAAGGGAAAGGGAAACCTGG	
LAMA1_358R	P17	TTCAATACCCAGGAAACTACTGTGG	312
SLC26A5_380F	P17	TGCTGTGGGTCATAACTTCCTG	
SLC26A5_380R	P17	AAAATTCCTTGTGAAGTAGGCAGTATC	464
DUSP13_205F	P18	ATGGACTCACTGCAGAAGCA	
DUSP13_205R	P18	CCACTCCTAACGTGGGTCTT	199
GPC6_198F	P18	CCTGCGTACGTCCTTGTGTA	
GPC6_198R	P18	GTGCCAGGAGAGTCTTCCAA	282
PDE2A_359F	P18	CAGGTTGAGATCCCCTGACC	
PDE2A_359R	P18	AGCCCTAGCCAGCCTCTCAG	578
SIAH3_443F	P18	TTGAACATCAGGGGAGTGGC	
SIAH3_443R	P18	TCCTGTTTCCTCAGCACCAAC	463
ADAMTS13_211F	P19	CTAATGGGGTCTGGCTCTTG	
ADAMTS13_211R	P19	CTGGTGAGCCTGGAAGACAT	270
APOB_206F	P19	AGGGAAAATCAAACACAGTGG	
APOB_206R	P19	AGTTGAGGGAGCCAGATTCA	396
CHPF_216F	P19	TCCAGAATACCAGCCATCTG	
CHPF_216R	P19	TGCAAGTCCAGCGTGTATTC	328
FBXO47_209F	P19	TAAAAGTGGTTCCCCGTGAG	
FBXO47_209R	P19	GCAGCTCCCTGTAGGTCAAT	214
FGF14_210F	P19	GGAAAACAGAAATGGGCAAA	
FGF14_210R	P19	GTTGACTGGTTTGCCTCCAT	228
IL3RA_213F	P19	AGGCGTCAACAGTACGAGTG	
IL3RA_213R	P19	AAATGTGAGAGGCCAACGTC	301
KRAS_214F	P19, P22	TTTGTATTAAAAGGTACTGGTGGAG	
KRAS_214R	P19, P22	CCTTTATCTGTATCAAAGAATGGTC	280
MYO18B_207F	P19	AAAACTGAGCTCTCCGACCA	
MYO18B_207R	P19	TGGGAGGAAGTCATCGAAGT	250
STX17_208F	P19	TGCTACACCTTCTATTGTGGAGA	
STX17_208R	P19	CTGTTGCTGCTGATGCTTCT	244
TKTL2_360F	P19	CCTTCAAAGGTCGGGGTATTC	
TKTL2_360R	P19	TTATCATTGACACCGTGGCG	781

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
C3orf15_225F	P20	GAGACTGCAGGAGGAGAGGA	
C3orf15_225R	P20	TTCTAGGCAGCTTCCACACA	178
CDK20_222F	P20	TAGTCCTAGACCCCCGAGGA	
CDK20_222R	P20	GAGCGACTGAGGGTGAGAAT	296
HOOK1_219F	P20	GATGTCTGGAAAATTGACCTTTG	
HOOK1_219R	P20	AGGCTTTTACAACCCCTCCC	156
OR9Q1_220F	P20	CTTTGGTTCCATCGACTGCT	
OR9Q1_220R	P20	AAGATCACCACCATGGAAGC	327
PCSK1_221F	P20	AAATATTGCAGAGCTGCCTGA	
PCSK1_221R	P20	GCTCCCTATGAAATTCTCCATC	339
SCEL_361aF	P20	CCCAGCCCTCTGAGATGAT	
SCEL_361aR	P20	GGGGGAGAAATAGAGAGTTCCA	227
SF3B1_218F	P20	GCTGCTGGTCTGGCTACTATG	
SF3B1_218R	P20	GAGTCCAGTCTGGGCAACAT	349
SF4_224F	P20	TCATCCTCATCCTCGTCCTC	
SF4_224R	P20	GAGATGCAAGGACAGCACAA	226
USP8_223F	P20	CAGCCAAGGCAAGGTTTTT	
USP8_223R	P20	GCATCTCAGGTTGGGTATTGA	378
CHRD_363F	P21	CAGCTGCCGCTGGTAAAGAC	
CHRD_363R	P21	TCATTCCCTCCCCAAAGGAC	582
CXCR4_239F	P21	TGAACCCCATCCTCTATGCT	
CXCR4_239R	P21	CATCTGTGTTAGCTGGAGTGAAA	178
DSP_231F	P21	GCTTCTTTCTTGGAATGTGAGG	
DSP_231R	P21	TTTCTGCAGGTTCTGATCCA	169
EYS_230F	P21	CTTGGGAGAAAGAATCTCTGTG	
EYS_230R	P21	GCCATCATAGTTTAGAGCCACA	181
FASTK_236F	P21	GAACTGACTCCCCACGTGAT	
FASTK_236R	P21	CCAAATACTCCACACGCAAA	183
HEATR3_362F	P21	TCAGTCCGGTTTTCACATGC	
HEATR3_362R	P21	TCCTACACCAACAGTTCTTCTCCC	288
MLH3_228F	P21	TGGTGTCATCCCAACATCAG	
MLH3_228R	P21	GAGGCTCTGATAAGAACATCTGA	177
SCN10A_233F	P21	TGCTGCAAACTGGATACCAC	
SCN10A_233R	P21	AGGGACTATGCCTCCCCTTA	156
TESK2_229F	P21	TCAAACCAGAAGAGAAGTAAAATTCA	
TESK2_229R	P21	CATTTCCTTCTCCTCCACCA	178
TP53_235F	P21	CATGAGCGCTGCTCAGATAG	
TP53_235R	P21	CCAAATACTCCACACGCAAA	166
TRAF3_234F	P21	GAAATAAAATAACGAGTGCTGGTG	
TRAF3_234R	P21	AGTCGCGAATCTTCCAGATG	231
VEGFC_227F	P21	GCTGTGGACCCCACAAAG	
VEGFC_227R	P21	ATTCACAGGCACATTTTCCA	177
ZDHHC20_232F	P21	TCTCATGAGTTGCCCTCACA	
ZDHHC20_232R	P21	TGCCTTCTTCAGCTCCATTC	223
AHSG_263F	P22	GTTGTCTTGCCTGGGAGGAG	
AHSG_263R	P22	AGGCTTGGACAAAATGGTGG	466
COL11A1_260F	P22	GCTTGAGTAGTAACAAAATCGCATC	
COL11A1_260R	P22	GAAAAGACATTAAGATGGACATGGAC	457
CSMD1_257F	P22	GTTTCACATGGCAATTCTTTGG	
CSMD1_257R	P22	CATATAAGCAAAATGGGAACATGTAAG	461
DDX3X_259F	P22	ATGCTGTGTTGAAAGCCCG	

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
DDX3X_259R	P22	TCCTAGCAAGTTCTTTCCTGCAC	595
DOCK4_256F	P22	TCTTTGATTTCCCCTTTTCTGTTAG	
DOCK4_256R	P22	AGCCACCTGCAACGTGACTG	284
FAT3_262F	P22	AACAACATCACGCTAGTGCAGG	
FAT3_262F	P22	TCCCGTAGCGTCCTCGC	457
KMO_264F	P22	TGTGCGTATCATCGTGTAGATGC	
KMO_264R	P22	TCAGTAATCCCAAAGTTTCAAAGC	346
SETD2_408F	P22	TTGCCAACAGTTTTGTATGGTTTAC	
SETD2_408R	P22	AATCAGAAAAGCCACCTCGC	258
TSHZ2_261F	P22	GCCAACATCCTGTCGGATTC	
TSHZ2_261R	P22	GAACACGGCCGATTGACATC	608
UCHL1_265F	P22	CAAGTCAGTTCAAGCACATTTCAC	
UCHL1_265R	P22	GCCAAGTGCCATAAAGTCACAC	256
WDR33_267F	P22	TCCCATCTACTTCTGGTCACACTG	
WDR33_267R	P22	GCAGCAGATACAGGCAAGGC	346
CACNB1_300F	P23	AACTTGCCTTGCTTTGTTGTG	
CACNB1_300R	P23	CTCGCTCCCTCCCCTCC	600
DDX43_410F	P23	GGAGAATGTCTTGCTTGGGG	
DDX43_410R	P23	TCTAACAATCAGATATCACTGGTTTCG	289
HERPUD2_365F	P23	TCCCTGTCCTGAAGTGGGAG	
HERPUD2_365R	P23	AACATTCTTTCATCATGTTGGTCC	344
MCC_301F	P23	CCCATCTGGCTGTTTTGCTC	
MCC_301R	P23	GAGCCTGGTGCTTCCAGATAAC	279
NRAS_299F	P23	ATGTGGCTCGCCAATTAACC	
NRAS_299R	P23	CAGAATATGGGTAAAGATGATCCGAC	240
PRMT1_307F	P23	CAGGACACGCTGTTCTCCAG	
PRMT1_307R	P23	GAAGCAGGGCATCACCCC	474
TBCC_364F	P23	CAGCGCGACGTTCTTTTGAC	
TBCC_364R	P23	TGTGACAATAAGTAAACTTCGAGACCC	525
UMOD_412F	P23	AGGGAAGGATCTCTGGGTGG	
UMOD_412R	P23	GGATCCGCGCACACGAG	499
MSH6_283F	P24	AACCTCTGCTTCCAGGTTCAAG	
MSH6_283R	P24	ATGGTGAGTGCGTGCTCTAAAA	579
MURC_281F	P24	CCTGTTGCCTGTTATCAAGCTG	
MURC_281R	P24	AAGATGTGACACTGGAAACCTCTG	587
ZNF292_282F	P24	CTCAACTGTGGAAGGCAGTGG	
ZNF292_282R	P24	TTCCCATTGGCTTTGGTCTG	663
C13orf26_366aF	P25	GACCTCTCCTCCCCTGTGTA	
C13orf26_366aR	P25	TGGCAAAAGCAATGTATTGTG	469
C1orf84_276F	P25	CTCCCATCCCTCAACCCC	
C1orf84_276R	P25	CTCTGCAAGGTTGGAGACCC	584
CDAN1_279F	P25	GTTTGCCTCCATTCCCTTCC	
CDAN1_279R	P25	GCAGCCCACTTCCTTTATTCC	540
CDC23_368F	P25	AAAGATAGTTCCTGGATGTTTGTTTG	
CDC23_368R	P25	AGTGGGTGACAATCCAACCC	505
ERC2_367F	P25	CAAGCGGCACTTGACATTTTC	
ERC2_367R	P25	GGATGTAGATATCTTCCAACCTGTG	583
FAT4_270F	P25	GAGATCTCTCTGCCTTTGTGGG	
FAT4_270R	P25	TGAGTGGCTTAGGGAGCTGG	582
HMCN1_272F	P25	TCTGTTTATCTCACAGGATGGCTG	
HMCN1_272R	P25	GGGACCCCAGTTATGTGTCC	502

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
ODZ1_275F	P25	GAAGACCTGGTGCTCATCGG	
ODZ1_275R	P25	TCCGTCACATTGAAAGGCAG	575
PKP2_274F	P25	CAGTATTTCTGGTCTCCTGGTTTG	
PKP2_274R	P25	GCATAAGCTGAGACCGAAGCC	420
POT1_271F	P25	TGGTTCGTAGGTTGTGCATCAG	
POT1_271R	P25	TGCATGAATATTGAGGCTCGTC	515
STC1_273F	P25	TGCCCATCCTCTTTGTCAGG	
STC1_273R	P25	CCCTCCCAGTCTGGCTCTG	348
VWF_277F	P25	TGAAAATGCCCAGACCAGTG	
VWF_277R	P25	CTCAGCCACAGCACCTCAGAT	520
ZC3H18_278F	P25	TGGAGCCAAGGCTTTCAGTG	
ZC3H18_278R	P25	AGAGCTGCATGCCGGACC	399

2.5.2 Oligonucleotides for Validation of Small Insertions and Deletions

Oligonucleotide	Patient	Sequence (5'→3')	Product Size (bp)
NOTCH1	P3	ACAGCTACTCCTCGCCTGTG	
NOTCH1	P3	TACTTGAAGGCCTCCGGAAT	205
DZIP1L_46aF	P4	GGATGCAAAGGGAATCTCG	
DZIP1L_46aR	P4	GCAAGAATGGGACCCTTACA	250
POLN_59F	P5	AATTCTGGAGAATAATGGAAAGTC	
POLN_59R	P5	ACCAGAAATGCCAAGGTCAC	473
RNF219_421F	P10	TTCAAACCATGTTGGCTTCAG	
RNF219_421R	P10	CCAGAGATTCTGTCCTTTCCTCAC	459
MBL2_134F	P12	GAAAGGTCAGTCCTGGGTCA	
MBL2_134R	P12	GGACTTTTTCCAGGGTCTCC	233
MYST4_135F	P12	CAGTAGGCAATCACCTGCAA	
MYST4_135R	P12	TTGGGGGAGAGCTTTGAATA	242
SPINK7_137F	P12	TGGAGAGTGTGCTGAATCTCAT	
SPINK7_137R	P12	CCTCTCGAATCCCCCTAAAC	250
SUSD4_429F	P16	CCGGACAGACTGGCTTATTG	
SUSD4_429R	P16	GCTAGGCTCTTTCTCCCCTC	288
MATN2_430F	P16	TCCTACCATTCCCTTTTCCTCTG	
MATN2_430R	P16	TGAGAGAGTTCCATGGTCACAG	263
OR4A16_424F	P21	CAGTCCCTGTTCTCCTGAGC	
OR4A16_424R	P21	GGGGCTGCCAATAGTAGTCA	207
CIB2_432F	P22	GGAAGAAAGTCTGGGAGCCG	
CIB2_432R	P22	GTGCCCCAGCCTCACCC	155
DNAJC2_434F	P22	GGGCAACAAGAGCAAAACTCTG	
DNAJC2_434R	P22	GCTGGGGACACCTCACTTTG	496
STAG2_405F	P22	CAGTGCCTCATTTATTGAACACC	
STAG2_405R	P22	TGTTAATTGAGATAGCACTGTAACTGG	440

2.6 Laboratory Equipment

Equipment	Company
Agilent 2100 Bioanalyzer	Agilent Technologies, Santa Clara, CA, USA
Balance BP 1200	Satorius, Göttingen, Germany
Blitz-Mikrozentrifuge	Fisher Scientific, Wohlen, Switzerland
cBOT	Illumina, San Diego, California, USA
Centrifuge 5415R	Eppendorf, Hamburg, Germany
Centrifuge Biofuge pico	Heraeus, Germany
Centrifuge Rotanta 460R	Hettich, Kirchlengern, Germany
Centrifuge Savant Speed Vac	Savant
Electrophoresis power supply	Gibco BRL by Life Technologies, USA
Fridge (+4 °C, -20°C)	Liebherr, Germany
Fridge Hera freeze (-80°C)	Heraeus, Germany
Gel Jet Imager	Intas
Genetic Analyzer automated DNA Sequencer ABI PRISM 3100	PE Applied Biosystems, Foster City, CA, USA
Glassware	Schott, Jena, Germany
Horizontal 11-14 Gelelectrophoresis System	Gibco BRL by Life Technologies, USA
Microwave	Panasonic
NanoDrop® ND 1000 spectrophotometer	Nano Drop Technologies, Wilmington, USA
Pippetes	Eppendorf, Hamburg, Germany
Qubit® Fluorometer	Invitrogen by Life Technologies GmbH, Darmstadt, Germany
Rotator	Labinco, Breda, Netherlands
Sonicator	Bioruptor Standard, Diagenode, Liège, Belgium
Thermocycler T Professional	Biometra, Göttingen, Germany
Thermocycler T3	Biometra, Göttingen, Germany
Thermomixer comfort	Eppendorf, Hamburg, Germany
Vortex-Genie II	Scientific Industries, Bohemia, N.Y., USA

2.7 Consumables

Consumables	Company
96 well plate septa (sequencing)	Applied Biosystems, Foster City, CA, USA
96 well sequencing plates for ABI 3100	Applied Biosystems, Foster City, CA, USA
Gloves latex	Semperit
Gloves nitril	Kimberly Clark
Cyclerseal Sealing Film	Axygen, Union City, CA, USA
Glassware	Schott, Jena, Germany
Microcentrifuge tubes 1.5 ml	Eppendorf, Hamburg, Germany
Non-skirted 96-well PCR plate 0.2 ml	Thermo Scientific, UK
NucleoFast 96 PCR plates (purification)	Macherey-Nagel, Dueren, Germany
Parafilm	Pechiney Plastic Packaging, Chicago, USA
PCR single cap 8er-soft strips 0.2 ml	Biozym, Oldendorf, Germany
PCR soft tubes 0.5 ml	Biozym, Oldendorf, Germany
PCR tubes 0.2 ml	Eppendorf, Hamburg, Germany
Pipette tips	Sarstedt, Nümbrecht, Germany
Qubit assay tubes 0.5 ml	Invitrogen by Life Technologies GmbH, Darmstadt, Germany
Strip caps (8)	Applied Biosystems, Foster City, CA, USA

2.8 Computer Operating System, Software and Programs

2.8.1 Software and Programs

Software	Link/ Company	Used for:
BEDtools	sourceforge.net	Coverage computation
BWA 0.5.8	bio-bwa.sourceforge.net	Mapping
Corel DRAW® 12	Corel	Cartoons / Figures
Data Collection Software 2.0	Applied Biosystems	Sequencing
EndNote X5	www.endnote.com	Bibliography management
Intas sotware	www.intas.de	Documentation of agarose gels
MS office	Microsoft	Text editing, data analysis, graph generation

NCBI database	www.ncbi.nlm.nih.gov	Multiple usage
Picard	picard.sourceforge.net	Summary statistics (exomes)
Primer 3 Plus	www.primer3plus.com	Primer design
R program	www.R-project.org.	Venn diagrams, heatmaps, boxplots
SAMtools	samtools.sourceforge.net	Analysis of exome data
Sequencing Analysis Software 5.1.1	Applied Biosystems	Base calling
snpEFF	snpeff.sourceforge.net	Variant Annotation
VarScan 2	varscan.sourceforge.net	Variant calling
Graph Pad Prism	GraphPad Software, Inc	Graphs

3 Methods and Patient Samples

3.1 REACH Study Cohort

The REACH study, conducted at 88 centers in 17 countries was performed in concordance with the principles of the Declaration of Helsinki based on ethic principles for medical research in human subjects. This trial was sponsored by F. Hoffmann-La Roche. All patients gave written informed consent after adequate explanation of aims, methods, objectives and potential hazards of the study (Roche Protocol BO17072F). This phase III trial enrolled patients with previously treated CLL to receive either Rituximab in combination with fludarabine and cyclophosphamide or fludarabine and cyclophosphamide alone. The study showed that fludarabine and cyclophosphamide in combination with Rituximab improved the outcome of patients with previously treated CLL (Robak *et al.* 2010).

3.2 DNA Extraction from Cell Pellets and Cell Lysates

Peripheral blood mononuclear cells were isolated routinely by Ficoll® density gradient centrifugation. After enrichment of lymphocytes, monocytes and macrophages, cell counts were obtained by a Sysmex Microcell Counter (Sysmex, Germany). Either aliquots of 5×10^6 cells were pelleted and frozen at -80 °C or the cells were lysed in 300 µl RLT buffer (Qiagen, Hilden, Germany; the exact composition is confidential) and stored at -80 °C.

DNA from frozen patient mononuclear cells was prepared with QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) for the purpose of screening, validation and exome sequencing. Pelleted cells were resuspended in PBS (2.4) to a final volume of 200 µl. Aliquots from RLT lysates were halved, one half was stored again at -80 °C and the other half was diluted with PBS to a final volume of 200 µl for further processing. Then 20 µl of Proteinase K was added and mixed well. For cell lysis, 200 µl of buffer AL was added and vigorously mixed on a vortexer and incubated for 10 minutes at 56 °C. AL is a guanidine based lysis buffer. For cells lysed in RLT buffer this step can be omitted, 200 µl of ethanol was added to the lysate and transferred to a QIAamp Mini spin column placed in a 2 ml collection tube and centrifuged for 1 minute at 8000 rpm. The flow through was discarded and the bound DNA was washed with 500 µl AW1 buffer and 500 µl AW2 buffer by centrifugation at 8000 rpm for 1 minute and 14000 rpm for 3 minutes, respectively. The flow through was discarded at each step. Thereafter the column was placed into a 1.5 ml microcentrifuge tube, 100 µl of AE buffer (2.4) was added and incubated at room temperature for 5 minutes. The DNA was eluted from the column by centrifugation at 8000 rpm for 1 minute and stored at -20 °C.

3.3 Shearing of DNA by Ultrasound

For the fragmentation of double stranded DNA (dsDNA) a sonicator (Bioruptor® Standard, Diagenode, Liège, Belgium) was employed, which sheares DNA by ultrasound at 20 kHz frequency. The principle of DNA shearing by ultrasound is based on the expansion and contraction of a liquid, when sound with a certain wavelength passes through it. During expansion, a cavity is formed which implodes at high energy and the resulting shearing forces fragment the DNA. For shearing, 3 µg of DNA was dissolved in 100 µl TE (2.4) buffer in a 0.5 µl hard plastic tube (Qubit assay tubes). Using a hard plastic tube is important to achieve the appropriate conditions for ultrasound transmission. Tubes were placed in the tube holder of the sonicator located in a water bath cooled with ice. The DNA was sheared at low power levels (160 W) for 3 x 15 minutes with an alternating 30 seconds “on” and 30 seconds “off” rhythm to prevent overheating of the samples. After 15 minutes of shearing, the ice in the water bath was changed to maintain the temperature. The tubes were centrifuged between ice changes to ensure that all the liquid was collected at the bottom of the tube.

3.4 Quantitative and Qualitative Analysis of Sheared DNA

3.4.1 Quantification of DNA by UV Spectrophotometry

Using UV spectrophotometry (Nano Drop® ND 1000 spectrophotometer), the concentration of DNA can be measured. Nucleic acids have their absorption maximum at 260 nm. The measured absorbance (A) can be used to calculate the DNA concentration using the Lambert-Beer equation. The absorbance of 280 nm of a DNA sample gives an estimation of protein contamination. The ratio of A[260]/A[280] reflects the purity of the sample, which is around 1.8 for pure DNA. As reference, buffer or ddH_2O was used depending on the solution the DNA was dissolved in. This method was mainly used to quantify DNA for validation or screening. With the Nano Drop instrument, dsDNA cannot be distinguished from single stranded DNA. Such a distinction is sometimes critical (see next paragraph).

3.4.2 Fluorescence-based Measurement of Double Stranded DNA Concentration

For exome sequencing it is important to know the exact concentration of dsDNA as the DNA is to be sheared to short dsDNA fragments for library construction. For this purpose, a fluorescence-based dsDNA quantification assay on a Qubit® Fluorometer (Life Technologies, GmbH, Darmstadt, Germany) was used. The fluorescent dyes are specific for dsDNA and become intensely fluorescent when bound to dsDNA. To measure the exact amount of dsDNA of a sample, the fluorometer is calibrated by two standards. The amount of dsDNA is directly proportional to the fluorescent signal. For each measurement, 2 µl of DNA solution was analyzed according to the manufacturer’s instructions.

3.4.3 Gel Electrophoresis

The quality of the DNA was also analyzed using agarose gel electrophoresis. DNA is a negatively charged molecule. In an electric field, DNA molecules move towards the anode and are separated according to their size when moving through a gel matrix. The concentration of the agarose gel was chosen depending on the DNA size. For fragments from 50 bp up to 3000 bp, 1.8 % (w/v) of agarose was dissolved in 1x TBE (2.4). Gels were prepared in an Erlenmeyer flask and boiled in a microwave oven until the solution was completely clear. After cooling to ≈50 °C, 10 µl GelRED™ from a 10,000 x stock solution was added to 100 ml of gel and transferred to a gel electrophoresis chamber. The gel is polymerized after approximately 20 min. Loading dye from a 5 x concentrated stock solution (2.4) was added to the samples before they were loaded on the gel. For sizing, a DNA ladder (2.4) was added. The DNA fragments were separated in the gel chamber containing 1x TBE electrophoresis buffer at 120 - 130 Volts (the distance between the electrodes was about 25 cm). After 20 - 60 min, the gels were analyzed under UV light. GelRED™ is a fluorescent dye, which interacts with the DNA backbone and can be seen under UV light. All gels were analyzed and photographed with the Intas Gel Get Imager.

3.4.4 Qualitative Analysis of DNA Libraries Using the Bioanalyzer

To determine the size distribution of small DNA samples, 1 µl sheared DNA or genome/exome library was analyzed on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The principle of separating charged DNA molecules by size in the Bioanalyzer is the same as described before using agarose gel electrophoresis. The major difference is that in the Bioanalyzer this is performed on a chip, which is a micro-channel based electrophoretic cell. Like an agarose gel electrophoresis, the chip is filled with a gel mixed with fluorescence dye. Electrodes provide an electric field, in which small molecules migrate faster than large ones through the gel matrix. The fluorescently labeled DNA is detected by a laser. In addition to the target DNA, a size standard with known DNA concentration is loaded in parallel. Additionally, two markers named “upper” and “lower” are loaded together with the sample DNA onto the chip as internal standards. The sample concentration is determined by the “upper” marker, which is of known concentration. The data are displayed as an electropherogram or as a gel-like image. The advantage of this method is the very small amount of DNA needed for sample characterization. The Agilent DNA 1000 Assay (Agilent Technologies, Santa Clara, CA, USA) analyzes DNA in the size range between 25 - 1000 bp down to a concentration of ≈0.1 ng/µl. The DNA 1000 chips were used for sheared DNA and adapter ligated libraries. Exome libraries were analyzed on an Agilent High Sensitivity chip (Agilent Technologies, Santa Clara, CA, USA), which is sensitive to concentrations between 5 and 500 pg/µl. For each measurement, 1 µl of sample (sheared DNA or library) was analyzed with a Bioanalyzer according to the user manual. This method is very well suited to determine the fragment size after DNA shearing and the shift in size after adapter ligation and amplification.

3.5 Enzymatic Manipulation of DNA

3.5.1 Polymerase Chain Reaction

PCR is a well-established technique to exponentially amplify segments of DNA flanked by two regions of known sequence. Synthetic oligonucleotides (primers) bind complementary to the 3' ends of a dsDNA fragment that is to be amplified. The first step of a PCR is the denaturation of dsDNA. Thereafter, the temperature is lowered for annealing so that the primers can bind to their targets. The annealing temperature is typically 2 - 4 °C below the melting temperature (T_m) of a primer, which can be calculated as follows: T_m = 4 °C x (G+C) + 2 °C x (A+T). The new strand is synthesized with a heat stable polymerase in the presence of dNTPs. A standard PCR reaction was performed using a PCR master mix (Taq PCR Master Mix Kit, Quiagen, Hilden, Germany) containing a *Taq* polymerase without proofreading activity.

PCR reagents	**Volume**	**Final Concentration**
Taq PCR Master Mix (2x):	12.5 µl	
5 units Taq DNA Polymerase		0.2 units Taq DNA Polymerase
2x PCR buffer		1x PCR buffer
400 µM of each dNTP		200 µM of each dNTP
3 mM $MgCl_2$		1.5 mM $MgCl_2$
10 µM Forward primer	1 µl	0.4 µM
10 µM Reverse primer	1 µl	0.4 µM
Q-Solution (5x) optional	5 µl	1x
25 mM $MgCl_2$ optional	1-2.5 µl	1 -2.5 mM
Template DNA 10 - 50 ng/µl	1 µl	-
ddH_2O	*ad* 25 µl	-

For GC rich fragments or fragments with a high degree of secondary structure for which the standard protocol failed, Q-solution or (Qiagen, Hilden, Germany) $MgCl_2$ was optionally added. The main component of Qiagen's Q-solution is betaine (N, N, N-trimethylglycine). This component changes the melting behavior of DNA in a way that GC and AT base pairs become equally stable (Rees *et al* 1993). If Q-Solution did not result in amplification, more $MgCl_2$ was added to a final concentration of up to 2.5 mM. $MgCl_2$ facilitates the binding between primers and template, which increases the chance to amplify difficult templates. The standard PCR protocol was cycled through a temperature program in a thermocycler as described below.

Steps	Temperature (°C)	Time (min:sec)	No. of cycles
Initial denaturation	95	05:00	1
Subsequent denaturations	95	00:30	30 - 35
Annealing	55 – 58	00:30	
Elongation	72	00:30	
Final Elongation	72	05:00	1

The annealing temperature depends on the melting temperature of the primers, and the elongation time is dependent on the product size and extension rate of the polymerase. The Qiagen Master Taq, which was most frequently used for DNA amplification, extends approximately 2 - 4 kb/min. For other polymerases the extension time was individually adjusted.

In many cases, a touchdown PCR was performed where the annealing temperature decreases from a high level down to its optimum. Starting with a high annealing temperature reduces unspecific products due to unspecific primer binding. The basic touchdown protocol was performed in a thermocycler as follows:

Steps	Temperature (°C)	Time (min:sec)	No. of cycles
Initial denaturation	95	05:00	1
Denaturation	95	00:30	2
Annealing	63 -2 °C per 2 cycles	00:30	
Elongation	72	01:00	
Denaturation	95	00:30	25 - 30
Annealing	58	00:30	
Elongation	72	00:30	
Final Elongation	72	05:00	1

3.5.1.1 *Purification of PCR Products by Ultrafiltration*

The resulting PCR products were cleaned on *NucleoFast 96 PCR Plates* (Macherey-Nagel, Düren, Germany) based on ultrafiltration technique. Each well contains a filter membrane through which the PCR fragments (≥150 bp) cannot pass, whereas primers, dNTPs and salts pass through the filter during centrifugation. For that procedure, PCR products were diluted with ddH_2O to a final volume of 100 µl and transferred to a NucleoFast® plate placed on top of a waste plate. The plates were centrifuged at 4000 x g for 20 minutes. Thereafter, the membrane was washed with 100 µl ddH_20 and centrifuged again under the same conditions. The purified PCR products were recovered by adding 25 µl ddH_2O or AE buffer onto the membrane. The purification plates were shaken for 10 minutes at room temperature before the PCR products were removed from the plates and stored at -20 °C.

3.5.1.2 *Purification of DNA and PCR Products using Magnetic Beads*

A polyethylene glycol (PEG) solution containing magnetic beads coated with a special solid phase is an alternative for PCR and DNA purification. PCR products and enzymatically treated DNA obtained during exome preparation are purified with Agencourt AMPure XP (Beckman Coulter GmbH, Krefeld, Germany) magnetic beads. DNA >100 bp is selectively bound to the coated paramagnetic particles. The concentration of salt and PEG in the bead solution determines the size of DNA bound to the beads. The higher the PEG concentration the smaller is the size of the bound DNA (Lis and Schleif 1975). DNA and magnetic beads are mixed at a ratio of 1:1.8 and incubated for 5 minutes. On a magnetic stand, the beads with the bound DNA are separated from contaminants, which remain in solution. After two washing steps with 500 µl of 70 % ethanol, the beads are dried at 37 °C in a thermoblock and resuspended in nuclease free water. The purified DNA can then be separated from the beads on a magnetic stand.

3.5.2 DNA Sequencing with Capillary Electrophoresis

The capillary sequencing method is based on the chain-determination method, which was developed by Sanger and colleagues in 1977 (Sanger *et al.* 1977). It is often called Sanger (dideoxy) sequencing. This method is based on the ability of the polymerase to incorporate nucleotides which lack the 3' hydroxyl group. Extension of DNA fragments is inhibited at the time a dideoxynucleotide is incorporated. For cycle sequencing, fluorescently labelled dideoxynucleotides are used. One reaction contains fluorescently labeled dideoxynucleotides (ddNTPs), template DNA, deoxynucleotides (dNTPs), a DNA polymerase, primer and buffer. For the cycle sequencing reaction, 1 - 2 µl purified PCR product, 0.3 µM primer, 2 µl 5x sequencing buffer (Applied Biosystems, Foster City, CA, USA) and 1.5 µl BigDye® Terminator v1.1 Ready Reaction mix (contains dNTPS, fluorescently labeled ddNTPS and a DNA Polymerase, Applied Biosystems, Foster City, CA, USA) are mixed in a final volume of 10 µl. The optimal extension activity of the polymerase provided by the kit is 60 °C. The reaction was cycled through a temperature profile in a thermocycler with the following program:

Steps	Temperature (°C)	Time (min:sec)	No. of cycles
Initial denaturation	94	03:00	1
Subsequent denaturation	94	00:30	25
Annealing	58	00:30	
Elongation	60	03:00	
Final Elongation	60	05:00	1

3.5.2.1 *Purifying Cycle Sequencing Products by Ethanol Precipitation*

Ethanol precipitation is a low cost method for purifying cycle sequencing products but it takes some time. The cycle sequencing products were acidified with 1/5 volume of 3 M NaAc and precipitated with 5 volumes of ethanol. After vigorous vortexing the precipitates were

centrifuged for 20 minutes at 4000 x g. The pellets were washed with 100 µl of 70 % ethanol and centrifuged again for 20 min at 4000 x g. After air drying, the pellets were resuspended in 20 - 30 µl ddH_2O. Thereafter, the purified cycle sequencing products were run on a capillary sequencer ABI PRISM 3100 (Applied Biosystems, Foster City, CA, USA). The resulting sequencing data was analyzed with the Sequencher (www.genecodes.com) software.

3.5.2.2 *Purifying Cycle Sequencing Products using Centri-Sep™ Columns*

Centri-Sep™ columns (Princeton Seperations, NJ, USA) were also used for purifying cycle sequencing products. This method is costly but very fast. The columns contain a hydrated polysaccharide matrix to remove excess dye terminators from the cycle sequencing reactions. The excess buffer of each column needs to be removed by an initial centrifugation step of 3 min at 750 x g. Thereafter, the cycle sequencing products can be loaded on the columns and unused material is removed in a second centrifugation step at 750 x g for 3 minutes. Excess dye terminators stay in the matrix and the cleaned products are directly collected in a fresh sequencing plate. The products can be directly loaded onto the sequencer.

3.5.3 Enzymatic DNA Manipulation for Exome Sample Preparation

The enzymatic manipulation of sheared DNA for exome sequencing was performed according to, the instructions of Agilent's (Agilent Technology, Santa Clara, California, USA) user guide: "SureSelect Target Enrichment System for Illumina Paired-End Sequencing Library" (Version 2.0.1, May 2010).

3.5.3.1 *Generating Blunt Ends and Phosphorylation of 5' Ends*

As a result of physical DNA shearing heterogeneous ends are produced, like overhanging 3 prime and 5 prime ends or blunt ends. Therefore, the fragmented DNA needs to be enzymatically treated to generate blunt ends for later adapter ligation. To generate blunt ends, the fragmented DNA is treated with Klenow enzyme (part of DNA Polymerase I) which fills up 5' overhangs with its 5'→3' polymerase activity in the presence of dNTPs and 3'overhangs are removed by its 3'→5' exonuclease activity. Additionally, T4 DNA polymerase is added to the mix which has the same properties but with a much higher exonuclease activity. The end repair mix also contains T4 polynucleotide kinase which catalyzes the transfer of a gamma phosphate from ATP to the 5' hydroxyl terminus of the fragmented DNA to ensure phosphorylated 5 prime ends, which are necessary for adaptor ligation later on. The end repair mix was prepared as follows:

Reagent	Volume for 1 exome library	Final concentration
DNA sample (100 ng/µl)	30 µl	30 ng/µl
10x Phosphorylation Reaction Buffer (NEB)	10 µl	1x
10mM each dNTP mix (NEB)	4 µl	0.4 mM
T4 DNA polymerase 3U/µl (NEB)	5 µl	0.15 U/µl

Reagent	Volume for 1 exome library	Final concentration
Klenow Fragment 5U/µl (NEB)	1 µl	0.05 U/µl
T4 PNK 10U/µl (NEB)	5 µl	0.5U/µl
Nuclease free water	*ad* 100 µl	-

The samples were incubated for 30 minutes at 20 °C in a thermocycler without closing the lid. Afterwards the blunted DNA was purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA). The DNA was eluted from the magnetic beads in ≈ 30 µl of nuclease free water.

3.5.3.2 *Adenylation of 3 Prime Ends*

The process in which one nucleotide A-overhangs are added to the 3' ends of blunted DNA is called tailing. In the following step the fragments can be ligated to Illumina's sequencing adapters as the sequencing primers have corresponding "T" overhangs on their 3 prime ends. Furthermore, A-tailing prevents the formation of concatamers of template DNA fragments during adapter ligation. A-tailing is performed using a Klenow fragment without 3'→5' exonuclease activity. For the reaction mix, 30 µl of end repaired DNA fragments, 5 µl of 10x NEBuffer2 for Klenow Fragment 3'→5' exo$^-$ (New England Biolabs, UK), 10 µl of 1 mM dATP (New England Biolabs, UK) and 3 µl of Klenow Fragment 3'→5' exo$^-$ were mixed together with nuclease free water to a final volume of 50 µl. The reaction mix was prepared as follows and incubated on a thermocycler at 37 °C for 30 minutes with open lid:

Reagent	Volumes for one exome library	Final concentration
End-repaired DNA fragments (100 ng/µl)	30 - 32 µl	60 ng/µl
10x NEBuffer2 for Klenow Fragment	5 µl	1x
1 mM dATP (NEB)	10 µl	0.2 mM
Klenow Fragment (3'→5' exo$^-$) (NEB)	3 µl	0.15 U/µl
Nuclease free water	*ad* 50 µl	-

The tailed DNA fragments were purified with magnetic beads and resuspended in nuclease free water.

3.5.3.3 *Adapter Ligation*

Adapter and DNA fragments are covalently linked during ligation. The molar ratio of adapters to DNA fragments was 10:1. Illumina provides forked adapters, which are ligated to the sticky ends of the target DNA. Due to the sticky and phosphorylated ends of the target DNA and the adapters, they are ligated in a manner that the resulting products have distinct sequences on their 3' and 5' ends shown in red and blue, respectively, in the illustration below. This is important, as in a subsequent PCR reaction (3.5.3.4) an anchor DNA sequence is added by using primers with a 5' extension. Figure 4 shows the adapter sequences (grey/blue and grey/red) and the genomic DNA fragment (black):

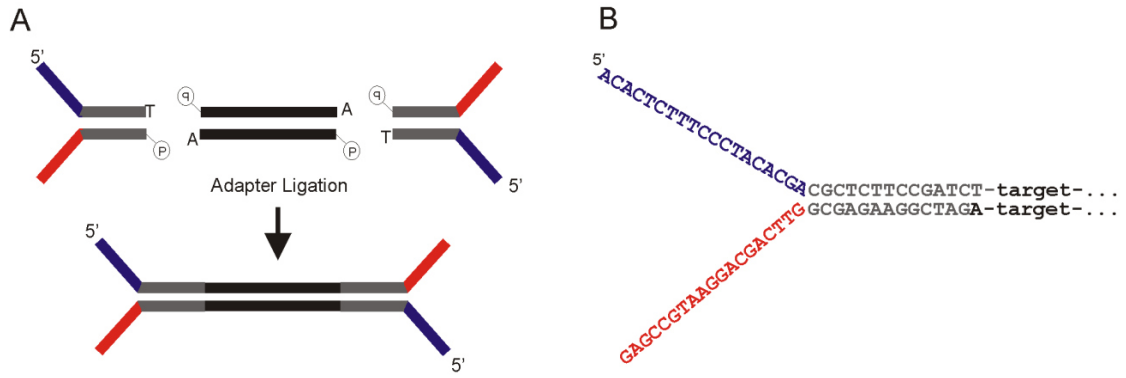


Figure 4 Illustration of Y-shaped adapters. A) The Y-shape of the adapters ensures that each fragment has different overhangs on both ends. B) Adapter sequence; Oligonucleotide sequences © 2007-2012 Illumina, Inc. All rights reserved.

For the adapter ligation, 14 µl of A-tailed DNA was mixed with 25 µl 2x DNA ligase buffer (New England Biolabs, UK), 6 µl paired-end Adapter oligo mix (Illumina, San Diego, California, USA) and 5 µl T4 DNA ligase (New England Biolabs, UK). The mix was incubated for 15 minutes at 20 °C on a thermocycler without using a lid. The ligated fragments were cleaned using magnetic beads and eluted in 50 µl nuclease free water. The adapters do not contain the anchor sequences for the flow cell. These sequences are introduced via PCR with specific primers binding to the overhangs of the adapters (the blue and red sequences in Figure 4).

3.5.3.4 *Amplifying Adapter-ligated Libraries*

After adapter ligation, a PCR was performed with Illumina primers PE 1.0 and PE 2.0. As mentioned earlier, these primers have specific 5' extensions so that the final amplification product can be bound and sequenced on a flow cell. The tails of the primers contain the P5 and P7 capture sequences, which bind to the complementary capture oligonucleotides on the flow cell.

The primer sequences as shown below were obtained from the "Illumina Adapters Sequences Letter" at www.illumina.com/support
(file name: 2012-09-18_IlluminaCustomerSequenceLetter.pdf).

PE PCR Primer 1.0
5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PE PCR Primer 2.0
5'CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Figure 5 Sequence of illumina primers. Oligonucleotide sequences © 2007-2012 Illumina, Inc. All rights reserved.

The sequence highlighted in orange and green are the P5 and P7 capture sequences, the adapter sequence is highlighted in blue/grey and red/grey with the terminal "T" in black. The underlined part represents the binding site of the sequencing primers.

During this PCR reaction, fragments with different adapters on either end are enriched in contrast to fragments containing only one or no adapter. Fragments with only one adapter can bind to the flow cell but cluster formation and sequencing is not possible. Fragments without any adapter cannot be bound to the flow cell at all. Besides fragments without adapters, the mixture can also contain adapter dimers. The terminal "T" on the primers only binds to fragments containing the corresponding "A" which was introduced earlier during the A-tailing process. In contrast, adapter dimers do not have the corresponding "A" and cannot be enriched. For that purpose, illumina designed special primers that are completely resistant to the 3'–5' exonuclease activity of the polymerase (illumina paired end sample preparation guide, Catalog # PE-930-1001, Part # 1005063 Rev. E, February 2011). Only a few PCR cycles are performed to enrich the adapter-ligated fragments in order to avoid introducing a bias for some genomic regions. The adapter-ligated fragments (25 µl) were amplified with 1 µl of Herculase II Fusion DNA Polymerase (this *Pfu*-based polymerase is manufactured by Agilent Technologies), 1 µl of each Illumina PCR primer (PE 1.0 and PE 2.0) with a final concentration of 0.25 µM, 0.5 µl of a 25 mM dNTP mix, 10 µl 5x Herculase II reaction buffer and nuclease free water in a 50 µl reaction. The PCR was performed in a thermocycler with an initial denaturation at 98 °C for 30 seconds, four cycles with 10 seconds denaturation at 98 °C, 30 seconds with 65 °C annealing temperature and 30 seconds elongation at 72 °C and a final elongation step for 5 minutes at 72 °C. In later experiments, we extended the initial temperature to 1 min and the denaturation step at each cycle was prolonged to 20 seconds, we also increased the number of cycles to six. The final genomic library contains fragments with P5 and P7 attachment sites on opposite ends. To examine if the adapter ligation was successful, the library was analyzed on a High Sensitivity Bioanalyzer chip. Typically, a size shift

from around 150 bp for the sheared DNA to a 250 - 300 bp peak for the adapter-ligated fragments was observed.

3.6 Exome Capture by In-Solution Hybridization

For exome enrichment, Agilent's Sure Select human all exon RNA library kit (V3) representing a target of 50 Mbp of coding sequences was used. This kit targets about ~1.6 % of the human genome (51,607,577 bases). The kit was developed by Agilent and the Sanger institute and its bait library was derived from different databases with an annotation based on the GENCODE project (Harrow *et al.* 2006). The bait library represents 99.86 % of CCDS (Sept.2009), 90.74 % of the GenBank (June 2010) sequence database, 96.47 % of RefSeq Genes, 97.50 % of RefSeq Transcripts (June 2010) and 75.24 % of Ensemble genes. In addition, miRNAs from miRBase are represented (taken from Agilent Technologies). The enrichment was performed according to the Sure Select target enrichment protocol (Version 2.0.1 May 2010).

The principle of enrichment is based on a system first described by Gnirke and colleagues 2009 (Gnirke *et al.* 2009). It is a solution-based oligonucleotide hybridization and capture method. Biotinylated RNA oligonucleotides, also called baits, are used for in-solution enrichment. The Sure Select 50 Mb kit contains 635,250 baits with a length of ~120 nucleotides representing 51.6 Mb of target sequence (Sulonen *et al.* 2011). The biotinylated baits are incubated together with the adapter ligated genomic library (see 1.5.7.4). The hybridized or captured fragments, representing the whole exome, were purified with streptavidin-coated magnetic beads.

The sequencing library (adapter ligated genomic fragments as described above), quantified using the Qubit (3.4.2), was dried in a vacuum centrifuge at 40 °C and dissolved in nuclease free water to a final concentration of 147 ng/µl. An aliquot of 3.4 µl containing 500 ng of total genomic library was used for the hybridization process. In parallel, buffers for hybridization were prepared. For the hybridization buffer, 25 µl of SureSelect Hyb #1 (contains 20x SSPE), 1 µl SureSelect Hyb #2 (contains 0.5 M EDTA), 10 µl SureSelect Hyb #3 (contains Denhardt's solution for blocking) and 13 µl SureSelect Hyb #4 (contains 10 % SDS) were mixed for each library. The hybridization buffer mix was kept at room temperature. In the meantime, the SureSelect block mixture was prepared from 2.5 µl SureSelect Block #1 (contains Human Cot-1DNA to block repetitive elements), 2.5 µl SureSelect Block #2 (contains salmon sperm to reduce non-specific hybridization) and 0.6 µl SureSelect Block #3 (contains an oligo mix to block the adapters). To prevent RNA baits from being degraded by RNases, 1 µl of SureSelect RNase Block was diluted in 2 µl nuclease free water. For each exome capture, 2 µl of freshly prepared RNase block was mixed with 5 µl capture library (bait) in a 0.2 ml PCR tube and put on ice. Each genomic library (3.4 µl) was mixed with 5.6 µl block mix in a 0.2 ml PCR tube on ice prior to hybridization, which was performed in a thermocycler using the following program.

Steps	Temperature (°C)	Time (h:min)
Preheating	95	Hold
Denaturation	95	00:05
Sample preparation	70	00:15
Hybridization	70 -1°C per h	05:00
Final hybridization	65	24:00

All following sample preparation steps were performed on a heated thermocycler. First the genomic library (3.4 µl library and 5.6 µl block mix) was placed in the thermocycler at 95 °C to allow denaturation of the genomic library. Thereafter, the thermocycler was cooled down to 70 °C, which is hot enough so that the library was not able to reanneal. Then 40 µl of hybridization buffer was placed in 0.2 ml PCR tubes in the thermocycler as well as the capture libraries. The lid of the thermocycler was closed to allow samples and buffers to reach the same temperature for 5 min. Then the lid of the thermocycler was opened and 13 µl of hybridization buffer was transferred to the capture library. Immediately after this, the entire genomic library (9 µl) was also transferred to the capture library and mixed by pipetting. The lid of the thermocycler was closed to start the hybridization process. The hybridization was performed at 65 °C for at least 24 h. The initial starting temperature of 70 °C was set to remove secondary structures from GC rich fragments. After the 24-hour hybridization step, the hot libraries were transferred to a solution of biotinylated beads (Dynal MyOne Streptavidin T1, Invitrogen) to separate the enriched, exonic DNA, which by now had annealed to the biotinylated baits, from the remaining genomic DNA.

3.6.1.1 *Enrichment of Captured Sequences using Biotinylated Beads*

The hybridized baits were enriched using streptavidin coated magnetic beads. The streptavidin-biotin interaction has one of the highest affinities of all non-covalent interactions. The washing solutions, elution and neutralization buffers were already included in the Agilent Sure Select human all exon RNA library kit (V3). For each hybridization, 50 µl of well-mixed Dynal magnetic beads were transferred to a 1.5 ml microzentrifuge tube. The beads were washed three times with 200 µl of Sure Select Binding buffer. Between the washing steps, the tube was put on a magnetic stand to remove the supernatant. The beads were resuspended in 200 µl binding buffer to which the captured libraries were transferred and mixed on a rotator for 20 minutes at room temperature. Thereafter, the solution was resuspended in 500 µl Sure Select Wash Buffer #1 and incubated at room temperature for 15 minutes. Beads and buffer were separated on the magnetic stand and the beads were washed three times with 500 µl prewarmed Sure Select Wash Buffer #2. For each washing step, the mix was incubated for 10 minutes at 65 °C in a thermoblock. For elution of the bound fragments, the beads were mixed with 50 µl Sure Select elution buffer and incubated for 10 minutes at room temperature. Beads and buffer were separated on a magnetic stand and the eluate was neutralized with 50 µl Sure Select neutralization buffer. The resulting exome library was concentrated from 100 µl to 30 µl using AMPure XP magnetic beads.

3.6.1.2 *Amplifying Captured DNA by PCR*

To increase the concentration of the captured library fragments, a few PCR cycles were performed. Only half of the library, 15 µl, was used for amplification and quantified by Qubit fluorometry. If the amount was too low, the other half of the library was amplified as well. In a 50 µl reaction, 15 µl of captured DNA was mixed with 10 µl Herculase reaction buffer, 1 µl of dNTPmix (25 mM each), 1 µl Herculase II Fusion DNA Polymerase, 1 µl Sure Select GA PCR Primers and nuclease free water. The PCR was performed on a thermocycler using the following program with optimized settings in parentheses.

Steps	Temperature (°C)	Time (min:sec)	Cycles
Preheating	98	02:00 (03:00)	1
Denaturation	98	00:20 (01:20)	
Annealing	60	00:30	10 - 12
Elongation	72	00:30	
Final elongation	72	05:00	1

Usually 10 PCR cycles were enough for adequate amplification. The quality and quantity of the library was analyzed on a Bioanalyzer using a High Sensitivity kit (3.4.4). If the concentration was too high and the chip was overloaded, the sample was either diluted and analyzed again or analyzed undiluted on a DNA 1000 kit. Alternatively, the molarity of a library can also be calculated from a Qubit concentration measurement if the molecular mass of the fragments is estimated. For this, the molecular mass for one base pair (≈ 650 g/mol) is multiplied with the median fragment length (250 - 300 bp). The molarity can be calculated by dividing the weight (g) by molecular mass (g/mol) and volume (L). A library concentration of around 2 ng/µl measured with Qubit Fluorometer equals 10 - 12 nM at a median fragment length of 250 – 300 bp.

3.7 Exome Sequencing on the Illumina Genome Analyzer IIx

The exome libraries were sequenced on a Genome Analyzer IIx (Illumina, San Diego, CA, USA). The sequencing was performed in collaboration with the Laboratory of Functional Genome Analysis (LaFuGa) at the Gene Center in Munich. The following paragraphs describe the basic principles of cluster generation and paired-end sequencing on an Illumina Genome Analyzer.

3.7.1 Cluster Generation on the cBot

The exome libraries were adjusted to 2 nM in 10 mM Tris-HCl, pH 8.5 buffer with 0.1 % Tween 20. The DNA was denaturated with 0.1 N NaOH in a 1:1 ratio and thereafter diluted with Illumina's hybridization buffer to a final concentration of 10 pM. The denaturated template DNA is now ready to be bound to the capture oligonucleotides of the flow cell. The flow cell is an optically transparent slide with 8 individual channels or lanes with covalently attached oligonucleotide anchors on the bottom surface of each lane. The loading of the target

fragments and bridge amplification is performed on a cBot (Illumina, SanDiego, CA, USA) device. The cBot instrument contains a thermal stage on which the flow cell is placed, as well as stage for the reagents. A manifold positioned on the flow cell and connected with the reagent stage provides the flow through of reagents and loading of template DNA fragments. On the cBot, clonal clusters are generated by bridge amplification as shown in Figure 6. The resulting clusters (≈50 million/lane) contain around 1000 identical copies of a single DNA fragment. For each exome, one lane was used. In one of the 8 lanes, a PhiX library was loaded as internal control. The flow cell is placed onto a thermal stage for cluster generation. The principle of cluster generation is shown below (Figure 6).

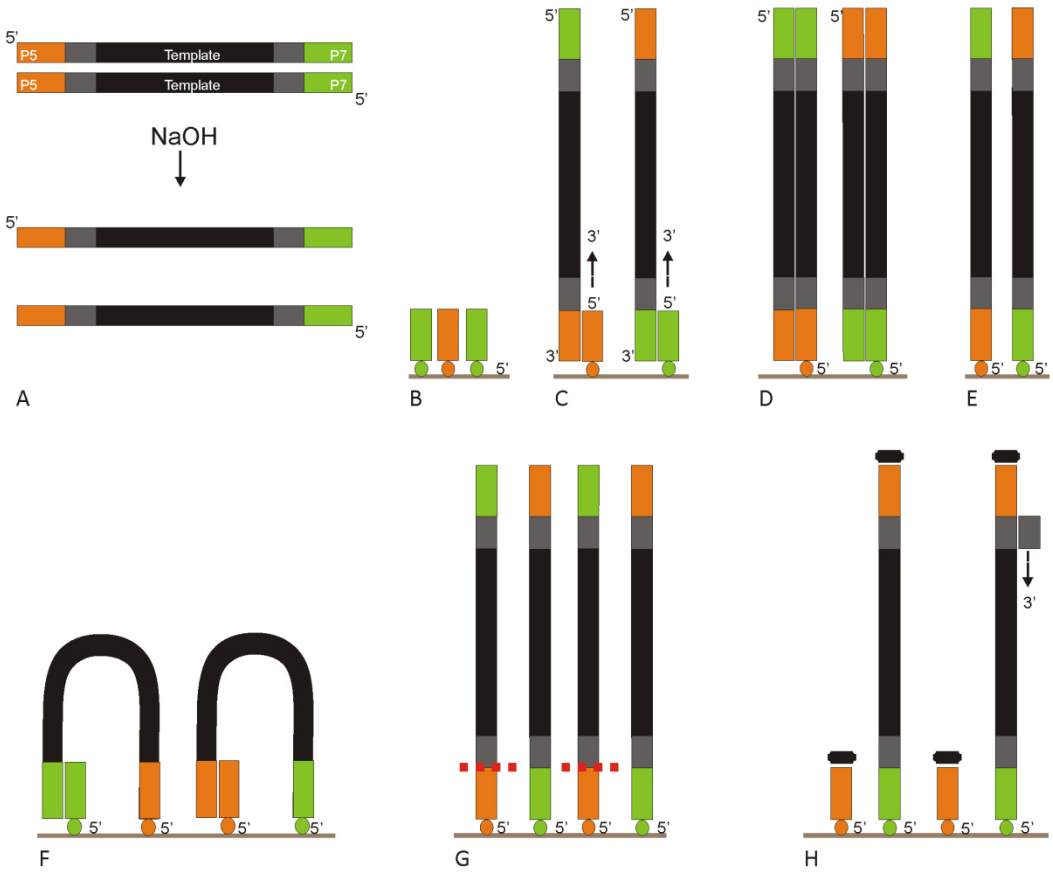


Figure 6 Schematic representation of cluster generation on the cBot. A) Double stranded templates with distinct adapter (P5 and P7) sequence on each end are denaturated prior to hybridization. B) The surface of the flow cell is coated with complementary adapters. C) Hybridization of fragments. D) Extension of bound molecules. E) The synthesized double strands are denaturated (formamide). The newly synthesized strand is covalently bound to the flow cell while the original strand is washed away. F) Single strands arch over to adjacent anchor molecules to form a bridge. After extension, the strands are separated. The whole process is repeated many times to form clonal clusters. G) Only one strand is sequenced in the first round. The other strand is cleaved enzymatically. H) The sequencing primers are bound and 3' ends are blocked to avoid unspecific binding during sequencing.

After amplification, the double stranded DNA with P7 (green) and P5 (orange) overhangs (a) is denaturated with NaOH to obtain single stranded fragments. On the cBot instrument, the single stranded library DNA fragments are immobilized on the flow cell through binding of the Illumina adapters to complementary anchor nucleotides (b), which are covalently attached to the flow cell. The fragments all bind with same orientation to the complementary oligonucleotides (c). The oligonucleotides are extended to generate identical copies of the fragments (d). The copied fragments are now covalently attached to the flow cell (e). The original strands are washed away after denaturation. During the next step clusters are generated through bridge amplification of each bound fragment. By arching over to adjacent anchor oligonucleotides, each template DNA forms a bridge like structure (f), which forms a priming site for the next extension step. The bridge is denaturated again. The process of isothermal denaturation (by formamide) and amplification is repeated 35 times to create a clonal cluster of each fragment with over 2000 molecules (illumina cBot™Part# 15006165 Rev. K., October 2012). After the last extension step, the fragments are linearized again and the fragments bound with the P5 adapters are enzymatically cleaved prior (g) to the binding of sequencing primers (h). Furthermore, the 3' ends are blocked (h) to avoid spurious DNA priming during sequencing. This whole process is fully automated and takes about 4 - 6 hours. After this, the flow cell is ready for sequencing on the genome analyzer.

3.7.2 Sequencing of Cluster-Amplified Template DNA

Illumina's sequencing is based on sequencing by synthesis using a mixture of four distinct fluorescently labeled dNTPs. Each nucleotide also has reversible terminator to cap the 3'-OH group. In each reaction all four nucleotide analogs are added but the DNA polymerase incorporates only the nucleotide which is complementary to the template DNA. Thereafter, the non-incorporated nucleotides are washed away, before the camera detects the fluorescent emission of the incorporated nucleotides. Before the next cycle starts, the fluorophore and the 3'-OH terminator are chemically removed.

For a more efficient use of the library we performed paired end sequencing. In paired end sequencing, each fragment is sequenced from both ends. The principle of sequencing is shown in Figure 7.

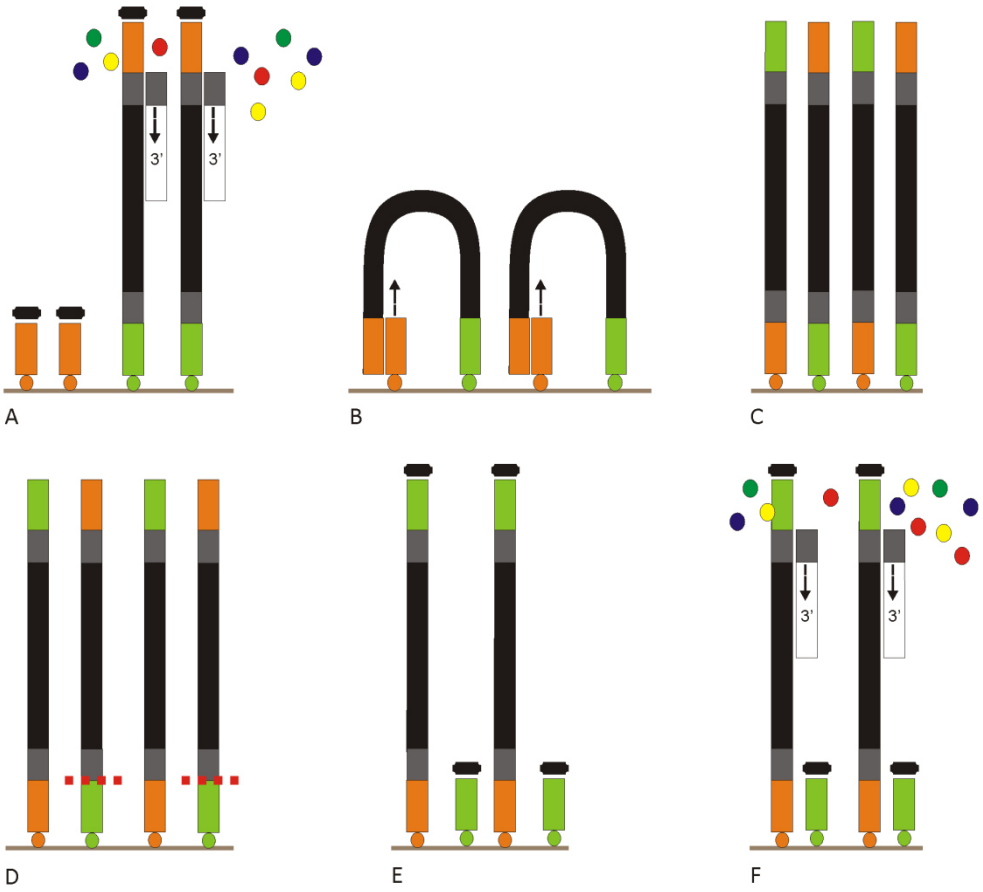


Figure 7 Illustration of paired end sequencing after cluster generation. A) Sequencing by synthesis with 4 fluorescent labeled nucleotides. As soon as the fluorescent terminators are incorporated (one per fragment) all clusters on the flow cell are imaged. Bound terminator and fluorescence is washed from the flow cell to start the next sequencing round. B) After the first read is completed, the strand synthesized by the sequencing reaction is removed and the covalently bound fragments arch over again to form bridges with adjacent anchors. C) Both anchor adapters on the flow cell are now covalently bound to the fragments. D) For the second read, all fragments with the P7 anchor are enzymatically cleaved. E) Blocking of 3' ends. F) The sequencing primer binds for the second read to the other end of the fragment.

After the cluster generation step on the cBot, the first of two sequencing rounds is started on the Genome Analyzer. Each cycle starts with the incorporation of a single, labeled nucleotide which has a blocking group at the 3'OH position (A). Excess enzyme and nucleotides are washed away and the fluorescence of the incorporated nucleotide is recorded with a camera. Then, the fluorescent dye and the blocking group are removed prior to the next sequencing cycle. After 76 - 80 cycles the first read is completed and the newly synthesized DNA strands are removed from the templates. The remaining covalently bound fragments form again bridges with anchor oligonucleotides and become extended (B). The resulting fragments are now attached to the flow cell in the opposite orientations (C). To ensure that all fragments are sequenced in the same direction, fragments are cleaved at the P7 adapter (D). After the 3' ends are blocked (E) and sequencing primers have bound, the fragments are sequenced from the other end (F). The pictures that are acquired after each incorporation step are processed

to obtain the raw sequence of each cluster. For a paired end run, about 14 days were required. This long run time is mainly due to the time it takes for the CCD camera to acquire the images. Each lane is imaged in 120 tiles with 4 different filters for each of the four fluorescent dyes used. Thus a single, one nucleotide sequencing step results in the acquisition of 4x120x8 = 3840 high resolution images.

3.8 Analyzing Exome Data on Galaxy Cluster

For large-scale data analysis an appropriate computer system is required. All data analysis was performed on a Unix computer cluster at the Gene Center accessible via a VPN client. The data was stored and processed on a Unix computer cluster (>5 nodes) with 25 terabyte of hard disk storage. We used Galaxy, an open-source, web-based application for NGS data analysis (Giardine *et al.* 2005; Goecks *et al.* 2010) locally installed and administered at the LaFuGa at the Gene Center in Munich. All open source based analysis tools used for analysis as described below were accessible from the Galaxy interface. This allows combining different analysis programs to generate individual workflows using a graphical user interface, instead of using the Unix command line interface.

3.8.1 Processing of Image Files and Base Calling

The raw images taken by the Genome Analyzer during a run contain all the sequencing information. The RTA1.9 software (www.illumina.com) performs the base calling on the primary image files in real time (while the run is in progress) and stores the resulting data in binary files. The raw images are deleted after base calling to conserve disk space. A single 76 bp run on a Genome Analyzer produces about 2.5 terabytes of primary image files. Therefore, the base calling cannot be repeated. The binary base calling data files are then converted into FASTQ files before being transferred to the Galaxy cluster. The FASTQ file format contains sequence and quality information.

```
@1:1:10629:1764:Y
AAGTGGAGCCCCCTCGCTGTACACCAGTGCGCTCACCATCGACGTATTAATGATGGGCCAGTGAGGCA
+1:1:10629:1764:Y
EGG@GGGGDGHHHGDGGGGFGGGGGHHDHGDHHHHG8EGGGGGGGGHHBFHFHHGDEGEFEBEBGEEF
```

Figure 8 Sample FASTQ file. This read originates from the first tile of lane one. The x and y coordinates define the cluster of this read, which is 10629:1764 in this example. In this case, the read passed the chastity filter (see Results) therefore it is flagged with Y (Yes). Reads that did not pass the chastity filter are marked with N (No). Line four describes the PHRED quality of each nucleotide by ASCII code.

A FASTQ file uses four lines per sequence, in the first line a sequence identifier and description of the read is given starting with the “@” character, followed by the sequence in line 2, in line 3 sequence identifier and description are optionally but the line always starts with the “+” character and line four contains the Phred quality value for each nucleotide in line 2 encoded in ASCII characters. In Figure 8 a read in FASTQ format from our own data is shown after base

calling and filtering with the CASAVA program. These FASTQ files are then used on the Galaxy server for downstream analyses.

3.8.2 Downstream Analysis of Sequence Data

3.8.2.1 *Burrows Wheeler Aligner (BWA)*

The first step in our downstream analysis of the sequence data was the mapping of the reads to the reference human genome. We used hg18 (NCBI Build 36.1/hg18) assembled in March 2006 by the International Human Genome Sequencing Consortium as the reference human genome sequence. This reference sequence has a total length of 3,104,054,490 bases.

There are many short read aligners available for the mapping of short paired-end reads based on different indexing strategies. Conventional alignment algorithms like BLAST are too computationally intensive to perform the alignment of millions of short reads to a reference genome in an acceptable time frame. The Burrows Wheeler Alignment tool is designed to rapidly map short reads to a reference sequence. It is based on the Burrows Wheeler transformation (Li and Durbin 2009). With this tool an index from the human genome DNA sequence is generated, comparable to an index of a book. With the use of the index, short reads can be mapped to the human genome. In contrast to other short read aligners, BWA allows mismatches and gaps. We used default settings for read mapping. The output format is SAM (Sequence Alignment/Map):

Column	Field	Description
1	**QNAME**	**Q**uery (pair) **NAME**
2	**FLAG**	bitwise **FLAG**
3	**RNAME**	**R**eference sequence **NAME**
4	**POS**	1-based leftmost mapping **POS**ition
5	**MAPQ**	**MAP**ping **Q**uality (Phred-scaled)
6	**CIGAR**	extended **CIGAR** string
7	**MRNM**	**M**ate **R**eference sequence **N**a**M**e ('=' if same as RNAME)
8	**MPOS**	1-based **M**ate **POS**ition
9	**ISIZE/TLEN**	**I**nferred insert **SIZE**/ Template **LEN**gth
10	**SEQ**	query **SEQ**uence on the same strand as the reference
11	**QUAL**	query **QUAL**ity (ASCII-33 gives the Phred base quality)
12	**OPT**	variable **OPT**ional fields in the format TAG:VTYPE:VALU

Table 1 SAM alignment format. The SAM format is a Tab delimited format and each line in the SAM format represents an alignment of a read. Each line consists of eleven mandatory fields and an additional optional field which is created by BWA. Description is taken from http://samtools.sourceforge.net/samtools.shtml.

The SAM format was developed in 2009 by Li and colleagues (Li *et al.* 2009). This format stores alignments from different sequencing platforms and read aligners in a generic file format, which can be used as input for further downstream analysis programs. By default, the output of BWA program is the alignment in a SAM file.

```
1:1:5215:1359:Y 83 Chr3 17028520 60 79M = 17028333 -266
GCCTTTCTGTGAGAAAAGGGAAGAAATCCAGGGAATATGCATCTTTGAGAACACTGTGGATTAAAACCGTGGATGAGGT
=67554?BB:@;EEEBF9E>AB@GBB>GHGGGEGBEGGHHHHEGGEGGGBHHHHHHHHHG@HHHHFGHHGGGGEBGGDG
XT:A:U NM:i:0 SM:i:37 AM:i:25 X0:i:1 X1:i:0 XM:i:0 XO:i:0 XG:i:0 MD:Z:79
```

Figure 9 Sample SAM alignment format.

It contains an optional header section (not shown) and an alignment section. The alignment section contains the information for each read (Figure 9). It tells us where the read maps to the reference genome and the quality of the read. All this information is given in the 11 mandatory fields of the alignment section.

FIELD	Example
QNAME	1:1:5215:1359:Y
FLAG	83
RNAME	Chr3
POS	17028520
MAPQ	60
CIGAR	79M
MRNM	=
MPOS	17028333
ISIZE/TLEN	-266
SEQ	GCCTTTCTGTGAGAAAAGGGAAGAAATCCAGGGAATATGCATCTTTGAGAACACTGTGGA…
QUAL	=67554?BB:@;EEEBF9E>AB@GBB>GHGGGEGBEGGHHHHEGGEGGGBHHHHHHHHHG…
OPT	XT:A:U NM:i:0 SM:i:37 AM:i:25 X0:i:1 X1:i:0 XM:i:0…

Table 2 Fields of the SAM format are assigned with an example of our own data.

In this example the cluster coordinates of a read on the flow cell is also the query name (QNAME). Each read of a pair has the same query name. A FLAG is used to describe the alignment, see also Table 3. In this example the RNAME is equivalent with the chromosome name where the read is located. The read we looking at, starts hat Chr3 position 17028520 with a mapping quality (MAPQ) of 60. CIGAR (compact idiosyncratic gapped alignment report) gives alignment information (see also http://samtools.sourceforge.net/SAMv1.pdf). The letter M in the CIGAR string denotes an alignment match which can be a sequence match or mismatch. In our example all 79 bases of the read align with the reference. The number of matches and mismatches is not given in this field. The equal sign denotes that the mate's reference sequence name (MRNM) is the same as the RNAME. The leftmost position of the mate is 17028333 (MPOS). The distance between the read pair calculated from the leftmost

position to the rightmost position is 266 bases (ISIZE/TLEN). The SAM format also contains the entire sequence (SEQ) of a read and the corresponding base quality in ASCII-33 format (QUAL). In the last field (OPT) there can be many optional tags (the NM tag for example gives the number of mismatches as an integer (i), (http://bio-bwa.sourceforge.net/bwa.shtml). In our example the number of mismatches is zero.

Bit	**Description** (http://samtools.sourceforge.net/samtools.shtml)	**Example** **FLAG: 83**
0x0001	the read is paired in sequencing	1
0x0002	the read is mapped in a proper pair	1
0x0004	the query sequence itself is unmapped	0
0x0008	the mate is unmapped	0
0x0010	strand of the query (1 for reverse)	1
0x0020	strand of the mate	0
0x0040	the read is the first read in a pair	1
0x0080	the read is the second read in a pair	
0x0100	the alignment is not primary	
0x0200	read did not pass quality controls	
0x0400	optical or PCR duplicate	

Table 3 Bitwise FLAG. Each bit has assigned information about the alignment. The FLAG in the example is 83 (decimal) in the binary system it would be 1010011. When reading the binary number from right to left the information assigned to each bit can be read from the table. In our example, the read has the following properties (indicated by 1): paired, mapped in a proper pair, read is from the reverse strand and it is the first read in a pair.

3.8.2.2 *SAMtools*

The SAM format stores alignments from different sequencing platforms and read aligners in a generic file format, which can be used as input for further downstream analysis programs. The SAM format is the human readable version, there is also a binary version called BAM which contains the same information. The SAMtools software package allows the manipulation of alignments in the SAM/BAM format (Li *et al.* 2009). The SAMtools software package was used to generate BAM files and sorted BAM files. In addition, it was used to remove PCR duplicates and to generate pileup and flagstat files. Flagstat files provide information on total mapped and paired reads for example. The pileup file describes the alignment information at each chromosomal position. The pileup, later mpileup (SAMtools 0.1.14), file was used as input for the variant caller software VarScan as described below.

SAMtools version 0.1.14 and higher includes a tool called BAQ: base alignment by quality. This tool helps to remove false positives variants close to an InDel. This option is turned on by default. As this default option is often too stringent, it was turned off for analyses and the less stringent variant calling routine integrated in SAMtools called extended BAQ was used instead. As later versions of SAMtools allow combining multiple SAM/BAM files into one pileup file, the

output is called mpileup instead. In addition, SAMtools provides a simple text alignment viewer, which allows visualizing reads mapped to a reference sequence.

3.8.2.3 *VarScan and VarScan 2*

Variant calling was performed using VarScan and VarScan 2 an open source tool (Koboldt *et al.* 2009; Koboldt *et al.* 2012). VarScan is specifically designed for short reads like the ones produced by the Illumina platform. Two different workflows were established for variant calling.

3.8.2.4 *Reference files*

As mentioned earlier, the output of each exome sequencing experiment was restricted to coding exons and known SNPs were excluded. This was achieved by working with BED (Browser Extensible Data) files containing the position of coding exons and SNPs. A BED file is a tab delimited text file, which is commonly used to represent genomic features. It has three required fields per line and nine optional fields (http://genome.ucsc.edu/). The first required field contains the chromosome name, the second and third lines contain start and end coordinates of a chromosome feature. To limit the sequencing data to known coding exons, a BED file from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) was downloaded in May 2012 containing all human RefSeq genes based on the human genome assembly from March 2006 (NCBI Build 36/hg18). To exclude known SNPs from the exome data the dbSNP build 130 BED file from the UCSC table browser was downloaded as well.

3.8.3 BEDtools

BEDtools is also an open source software tool, which allows manipulating and comparing genomic features contained in different BED files (Quinlan and Hall 2010). We used BEDtools to perform coverage calculations across the whole exome and for individual genes.

3.8.4 Picard

Summary statistics were calculated with the Picard program (http://picard.sourceforge.net). BAM files were analyzed for general alignment metrics (summary alignment metrics) and exome hybridization metrics (hybrid selection specific metrics).

4 Results

4.1 Whole Exome Sequencing of 25 CLL Patient Samples

To identify disease-causing mutations in malignant diseases, next-generation sequencing is the method of choice. The terms next-generation sequencing or second-generation sequencing refers to techniques and machines that allow the generation of mega- or gigabases of sequence information in a single experiment. All next-generation sequencing methods are based on the massive parallelization and miniaturization of sequence reactions. Several sequencing systems are commercially available and include the 454 pyrosequencing technology (Roche) developed by Jonathan Rothberg (454 Life Sciences), ligation based sequencing in the SOLID system (ABI) or Illumina's sequencing by synthesis strategy. With constant improvements in these technologies, sequencing costs per megabase are decreasing tremendously. With the latest next-generation sequencing technology whole genomes or targeted regions can be analyzed and either mapped to a reference or assembled *de novo* in a very short time. However, whole human genome sequencing is still costly and requires the manipulation of substantial data quantities. For many problems the sequencing and the analysis of the coding regions of a genome, the so-called exome, is quite sufficient to gain insight into disease pathology. To perform whole exome sequencing (WES) one can choose between different exome enrichment strategies.

We performed WES of 25 CLL patient samples and their matched normal controls using an in solution exon enrichment strategy (Agilent SureSelect). In CLL several prognostic markers, like the deletion of certain genomic regions as assayed by FISH, the mutational status of the *IGHV* locus or the expression levels of *ZAP-70* have been identified. However, the predictive power of these markers is limited and they cannot fully explain the heterogeneity and the great variability in the clinical course of disease seen in CLL patients. The use of the monoclonal anti-CD20 antibody (Rituximab) in conjunction with chemotherapy has significantly improved progression-free survival of CLL patients even in second line therapy (Robak *et al.* 2010). The aim of this project was to identify new, potentially predictive markers in previously treated patients that had reached complete remission after second-line treatment with fludarabine and cyclophosphamide (FCR) or with fludarabine and cyclophosphamide (FC) alone.

4.1.1 Biological and Clinical Characteristics of the CLL Patient Cohort

The CLL patient samples used in this project were obtained from the REACH study patient cohort. The primary objective of the REACH study was to compare progression free survival in previously treated CLL patients, treated with rituximab and fludarabine and cyclophosphamide (FCR) with patients treated only with fludarabine and cyclophosphamide (FC) (Clinical Study Protocol BO17072F). This study was an international, multicenter, open-label, phase III trial. It was conducted at 88 centres. Between July 2003 and August 2007, 552 patients were

randomized to the FC or the FCR arm (Robak *et al.* 2010). The Laboratory for Leukemia Diagnostics at the university hospital Grosshadern in Munich performed the routine diagnostics for these samples. The routine diagnostic tests included morphology, FISH analysis using a panel of 5 FISH probes, measurement of ZAP-70 expression levels and determination of the *IGHV* mutational status.

The CLL patient samples for WES were selected from the REACH study cohort (n=552). Patient samples were selected based on the availability of a remission sample from the same patient and a negative minimal residual disease (MRD) status. In total, 25 CLL patient samples and their matched normal controls fulfilled these criteria. Table 4 shows a comparison of the clinical and biological characteristics of the patients available for WES with the complete REACH study cohort. Note that the data of the 25 WES patients are included in the REACH cohort data.

Characteristics	**REACH cohort**	**WES subgroup**	**P-value***
Number of patients, n			
Sex, n	546	25	
Female	181 (33 %)	8 (32 %)	1.0
Male	365 (67 %)	17 (68 %)	
Age, n	546	25	
Mean (sd)	61.8 (9.0)	62.1 (9.8)	0.85
Median (range)	63 (35-83)	61 (38-79)	
Binet stage, n	546	25	**0.023**
A	55 (10 %)	6 (24 %)	
B	321 (59 %)	16 (64 %)	
C	170 (31 %)	3 (12 %)	
Beta-2 microglobulin (mg/L), n	531	25	0.35
Mean (sd)	3.4 (2.2)	3.41 (1.3)	
Median (range)	3.47 (0.003-17.1)	3.0 (1.8-6.5)	
Lymphocytes (10^9/L), n	544	25	0.26
Mean (sd)	64.5 (68.3)	45.5 (41.6)	
Median (range)	42.0 (0.54-431.0)	25.4 (5.2-148.0)	
B-Symptoms, n	546	25	0.66
No	391 (72 %)	17 (68 %)	
Yes	155 (28 %)	8 (32 %)	
ECOG performance status, n	545	25	1.0
0	325 (60 %)	15 (60 %)	
1	220 (40 %)	10 (40 %)	

Characteristics Number of patients, n	REACH cohort	WES subgroup	P-value*
***IGHV* mutational status, n**	517	25	**0.033**
Mutated	191 (37 %)	15 (60 %)	
Unmutated	326 (63 %)	10 (40 %)	
***ZAP-70*, n**	408	21	**0.04**
Negative	236 (58 %)	17 (81 %)	
Positive	172 (42 %)	4 (19 %)	
***CD38*, n**	324	16	0.31
Negative	154 (47.5 %)	10 (62.5 %)	
Positive	170 (52.5 %)	6 (37.5 %)	
Del17p, n	529	25	0.25
No	487 (92 %)	25 (100 %)	
Yes	42 (8 %)	0 (0 %)	
Del13q, n	531	25	0.68
No	225 (42 %)	12 (48 %)	
Yes	306 (58 %)	13 (52 %)	
Trisomy 12, n	530	25	0.55
No	462 (87 %)	21 (84 %)	
Yes	68 (13 %)	4 (16 %)	
Previous Chemotherapy, n	537	25	0.82
Alkylator refractory	145 (27 %)	6 (24 %)	
Alkylator sensitive	302 (56 %)	16 (64 %)	
Fludarabine	90 (17 %)	3 (12 %)	
Treatment at 2nd line, n	546	25	0.22
FCR	272 (50 %)	9 (36 %)	
FC	274 (50 %)	16 (64 %)	
Time from last progression (days), n	544	25	0.17
Mean (sd)	98.8 (145.4)	53.6 (42.4)	
Median (range)	50 (1-1406)	41 (4-171)	

Table 4 Clinical and biological characteristics. Clinical and biological characteristics of the WES patients (n=25) compared with the total REACH cohort. These data were kindly provided by F. Hoffmann-La Roche. There are significant differences in *ZAP-70* status, *IGHV* mutational status and Binet stage between both cohorts. * P-values were calculated for various clinical and biological variables applying the Mann–Whitney U test or Fisher's exact test respectively for continuous and categorical variables (*, P <0.05).

The fact that the REACH study included only patients at relapse is the reason for the adverse clinical and biological characteristics of this cohort. Most of the REACH patients had an intermediate Binet stage at diagnosis. More than 60 % of the patients had an unmutated *IGHV* status, which is associated with poor prognosis (Damle *et al.* 1999; Hamblin *et al.* 1999). The WES subgroup of patients, in contrast, has more favorable prognostic markers. The WES patients had significantly better *IGHV* status, 60 % had a mutated and only 40 % an unmutated status. There was also a significant difference in the expression levels of *ZAP-70*, a surrogate marker of *IGHV* mutational status, with most WES patients being *ZAP-70* negative. According to the Binet staging system patients selected for whole exome sequencing had a significantly better prognosis compared to the entire REACH study cohort. Obviously there is a selection bias in the exome cohort towards patients with better prognosis.

4.1.1.1 *FISH Status and Content of CD19 Positive Cells in Tumor Samples*

The deletion of certain genomic regions as assayed by FISH is one of several markers of prognostic relevance in CLL. Patients harboring a 13q deletion as sole abnormality have the most favorable prognosis. In contrast patients harboring a 17p deletion have the worst prognosis, as the *TP53* locus is affected. Patients with 11q deletions affecting the *ATM* locus have a slightly better prognosis. Patients with no abnormality in the FISH screen or with a trisomy 12 have an intermediate prognosis (Dohner *et al.* 2000). 20 of the 25 WES samples had a single genomic alteration when analyzed with FISH probes for the following genomic regions: 6q, 13q, 11q, 17p and centromere 12 (for the detection of trisomy 12) (Table 5).

	del (13q)	**No aberration**	**Trisomy 12**	**del (11q)**	**del (6q)**
Samples (#)	13	5	4	2	1
Samples (%)	52	20	16	8	4

Table 5 Chromosomal aberrations in REACH exome cohort. The most frequent abnormality in the REACH exome cohort is the 13q deletion in n=13 patients, followed by normal karyotype (n=5), trisomy 12 (n=4), 11q (n=2) and 6q (n=1).

Most patients in the WES group harbored a 13q deletion as sole aberration (52 %). There was no significant difference in the proportion of patients with 13q deletion compared to the total REACH cohort. No FISH aberration, trisomy 12, 11q deletion and 6q deletion were found in 20, 16, 8 and 4 percent of the WES patients, respectively. There were no WES patients with a deletion of the short arm of chromosome 17. This is due to the fact that the WES patients had to have a remission sample. Patients with a 17p deletion affecting the *TP53* tumor suppressor gene have a low chance of achieving remission.

	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25
IGVH mutated	■	■			■		■	■	■			■	■			■	■	■	■		■			■	■
IGVH unmutated			■	■		■				■	■			■	■					■		■	■		
no FISH marker		■	■																■	■	■				
Trisomy 12					■	■																	■	■	
del(13q)	■			■			■	■	■			■	■	■	■	■	■	■							■
del(11q)											■											■			
del(17p)																									
del(6q)										■															
CD19 positive cells (%)	89.50	NA	88.53	80.95	93.50	84.05	88.69	90.13	79.83	69.42	77.05	53.99	72.02	84.39	93.14	87.80	86.56	NA	93.98	69.31	70.53	92.84	48.59	82.99	71.87

Table 6 Overview of FISH markers and percentage of CD19 positive cells as measured by flow cytometry for each tumor sample. For patients P02 and P18 no flow cytometry data was available (NA = not analyzed).

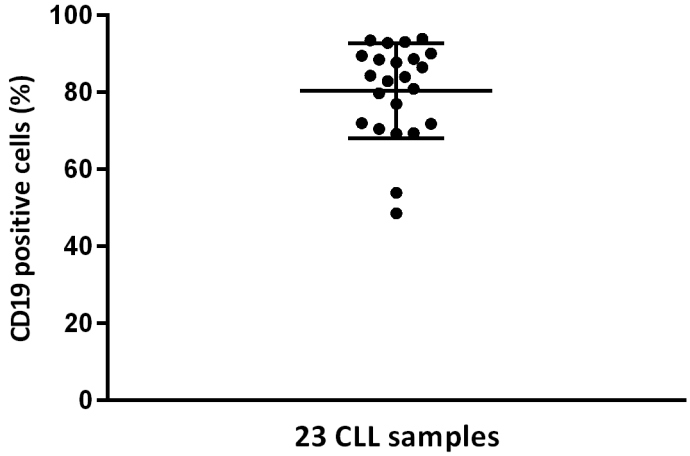


Figure 10 Proportion of CD19 positive cells as measured by flow cytometry. For 23/25 samples data was available at diagnosis.

To estimate the proportion of CLL cells in the leukemic samples derived from peripheral blood, routinely obtained flow cytometry data were analyzed. Flow cytometry data was available for 23 of the 25 WES samples. As B cell derived cells, most B-CLL cells express the CD19 antigen on their cell membrane. The proportion of CD19 positive cells in our WES samples ranged from 49 to 94 %. As CD19 is also found on normal B-cells the exact proportion of CLL cells cannot be determined. An enrichment of CD19 positive cells was not performed.

4.1.1.2 *Disease Free Remission Samples as Defined by MRD and FISH Status*

The 25 CLL samples and matched remission samples for WES were selected from the REACH study cohort. As we used the remission sample from the same patient as non-tumor control, we had to ensure that there were only very low levels of residual CLL cells in the remission sample. All non-tumor samples for which sufficient material was available (>5 µg of gDNA) were selected according to MRD level and FISH status. The remission status was determined by measuring the level of the patient-specific *IGHV* rearrangement using realtime quantitative PCR. These measurements were performed routinely for all samples of the REACH study cohort. Disease free remission was defined as MRD levels of less than $1x10^{-3}$ as measured using the disease-specific *IGHV* rearrangements.

Remission Sample	MRD Level	MRD Sensitivity	MRD Quantitative Range	MRD Interpretation
P00	$<5x10^{-5}$	$5x10^{-5}$	$1x10^{-4}$	Negative
P01	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P02	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P03	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P06	$<1x10^{-3}$	$1x10^{-4}$	$1x10^{-3}$	Oqrpos
P07	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P08	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P09	$4.75x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P12	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P13	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P16	$7.7x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P19	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-3}$	Negative
P20	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P21	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-4}$	Negative
P23	$<5x10^{-5}$	$5x10^{-5}$	$1x10^{-4}$	Negative
P24	$<1x10^{-4}$	$1x10^{-4}$	$1x10^{-3}$	Negative

Table 7 MRD levels as determined by quantitative realtime PCR. MRD levels were routinely measured using disease specific *IGHV* rearrangements by realtime quantitative PCR. Disease free remission was defined by MRD levels of less than $1x10^{-3}$. Oqrpos: out of quantitative range positive; iqrpos: in quantitative range positive.

MRD levels of 16 of the 25 patients of the WES cohort had MRD levels of less than $1x10^{-3}$. Of these 16 patients n=2 had an MRD level of $<5x10^{-5}$, n=11 had an MRD level of $<1x10^{-4}$ and n=3 had an MRD level of $<1x10^{-3}$. To increase the number of samples available for WES we also included samples with a FISH negative remission sample. FISH screening was routinely performed on all patient samples. We estimate that the maximum proportion of remaining CLL cells in a remission sample is no more than 10 % if the FISH abnormality present at diagnosis can no longer be detected. Nine additional samples from the REACH cohort fulfilled this requirement (Table 8).

Remission Sample	MRD Level	MRD Sensitivity	MRD Quantitative Range	MRD Interpretation
P04	$1.25x10^{-3}$	$1x10^{-4}$	$1x10^{-3}$	Iqrpos
P05	$7.44x10^{-3}$	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P10	$6.09x10^{-3}$	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P11	$1.16x10^{-3}$	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P14	0.02	$1x10^{-4}$	$5x10^{-4}$	Iqrpos

Remission Sample	MRD Level	MRD Sensitivity	MRD Quantitative Range	MRD Interpretation
P15	0.02	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P17	0.04	$5x10^{-5}$	$5x10^{-5}$	Iqrpos
P18	0.31	$1x10^{-4}$	$1x10^{-4}$	Iqrpos
P25	0.02	$1x10^{-4}$	$1x10^{-4}$	Iqrpos

Table 8 MRD levels of the samples selected on the basis of a negative FISH marker in the remission sample (n=9). MRD levels were routinely measured by realtime quantitative PCR. Iqrpos: in quantitative range positive.

Out of these nine samples, n=4 had fewer than 0.74 % CLL cells according to MRD measurements, n=4 had fewer than 4 % CLL cells. There is only one remission sample, which was FISH negative but had an MRD level of 0.31 (30 %).

4.1.2 Target Enrichment, Sequencing and Low Level Data Analysis Including Read Mapping and Quality Metrics

We used the Agilent SureSelect Human All Exon 50 Mb platform (later referred to as Agilent 50 Mb kit) to enrich the coding regions, or exome, of each CLL sample. The Agilent 50 Mb kit consists of a biotinylated RNA library of fragments with a length of 120 nucleotides, the so-called baits, which are complementary to the coding regions of the human genome. The enrichment is based on an in-solution hybridization step. To enrich the patients' exome, genomic DNA is converted into a sequencing library, which is composed of fragments of about 150 bp in length. Then, those fragments from the genomic library are enriched that contain exonic sequences (3.6). This process takes at least 24 h. The enriched fragments are amplified, purified and diluted to an appropriate concentration for sequencing. The exome libraries were sequenced on a Genome Analyzer IIx (Illumina) and the data were analyzed on a Galaxy server. The sequencing and the data analysis were performed at the Laboratory for Functional Genome Analysis (LaFuGa) at the Gene Center of the University of Munich.

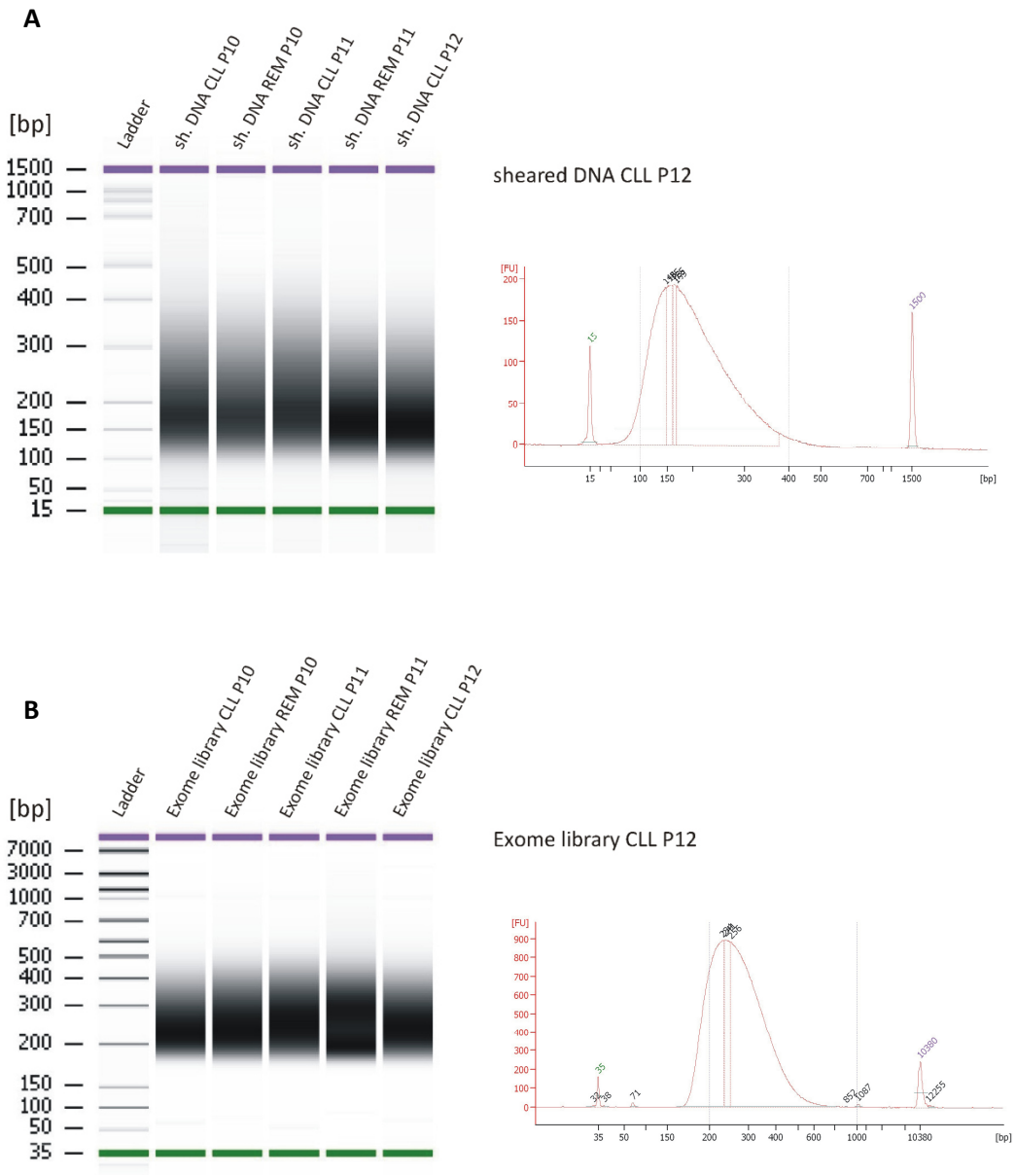


Figure 11 Gel like images and electropherograms of two Bioanalyzer runs. A) Sheared DNAs from five different samples run on a Bioanalyzer DNA 1000 chip with fragments of around 150 bp in length represented as gel like pictures (left) and as electropherogram (right) of one patient sample (P12). The amount of DNA is calculated from the area below the curve. B) The second gel like picture represents the same patient samples but after adapter ligation and post-hybridization amplification. There is a clear shift visible from 150 bp to 250 bp. The shift can be also seen in the electropherogram. In addition, there is a small peak visible at around 71 bp representing excess primers. Concatamers can be seen at around 1000 bp.

Figure 11 shows the results after DNA shearing of genomic DNA (A) and the shift after adapter ligation and amplification (B).

4.1.2.1 *Low Level Data Analysis*

The base calling from the raw image files and the generation of the FASTQ files was performed by the Genome Analyzer IIx during the sequencing run. Thereafter FASTQ files were transferred to the Galaxy server. The low level data analysis on the galaxy server is summarized in Figure 12, briefly: at the beginning, reads that did not pass illumina's chastity filter need to be removed, in the following the reads can be mapped to the reference genome and a generic file is generated which is the basis for further downstream analysis including SNV-, InDel-calling and calculation of coverage.

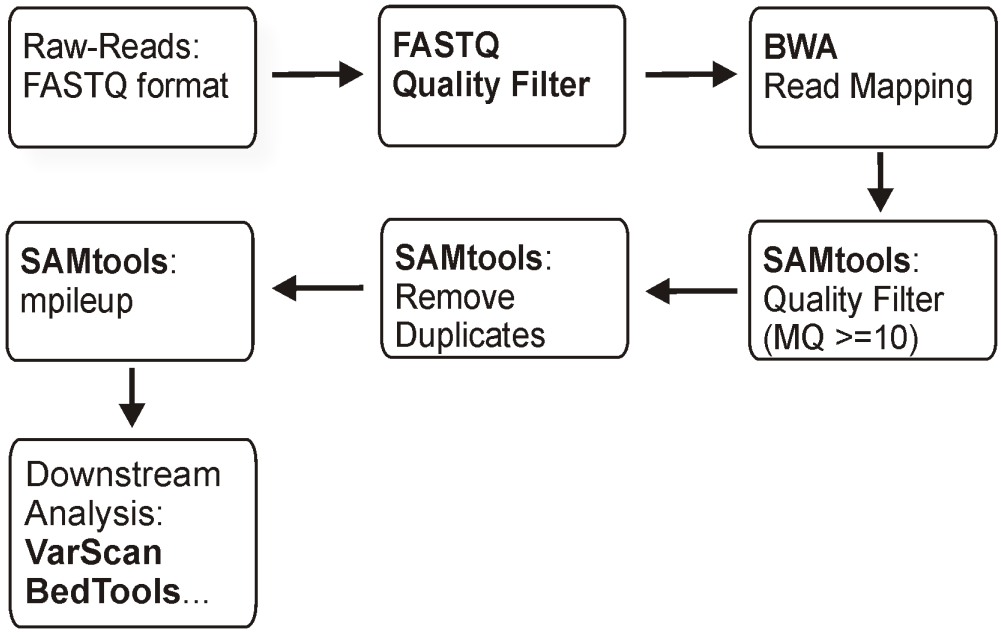


Figure 12 Flowchart of the initial sequence analysis steps. The images acquired during the sequencing run were processed in real time on the Genome Analyzer (cluster detection, base calling etc.) and FASTQ files were generated, which were transferred to the Galaxy server. With the FASTQ Quality Filter, reads that did not pass chastity filter were removed. Reads were mapped to the human reference genome with BWA default settings. SAMtools was used to remove duplicated reads and reads with a mapping quality of <10. To generate the mpileup file, which is the input format for downstream analysis, the minimum base quality for a base to be considered was set to 13 and the "extended BAQ computation" option was selected.

The chastity filter identifies clusters with a low signal to noise ratio. Frequently, adjacent clusters are too close to each other or even overlapping and their signals cannot be measured independently.

The chastity of a base call is defined as: $chastity = \frac{brightest\ intenstiy\ (\text{that of the base itself})}{brightest+second\ brightest\ intensity}$

A cluster passes the chastity filter if not more than one base call has a chastity of <0.6 in the 25 first bases of a cluster. A chastity of ≥0.6 means that the signal of the base is at least 1.5 fold stronger than the signal of the second most intense base at that position. The chastity filter information is encoded as comment in the first line of each FASTQ file. Reads are either marked with "Y" (Yes) or "N" (No) for reads that passed or failed the chastity filter, respectively. As reads that did not pass the filter are only flagged and not deleted, we used the

'FASTQ Quality Filter' an in house tool, to remove all reads which did not pass the chastity filter. At this point, the reads can be mapped to the reference genome. Mapping is a challenging and very time intensive step of NGS data analysis. An enormous number of short reads need to be mapped to a reference genome to create an alignment file. We decided to use the BWA mapping tool, its algorithm is designed to align short reads up to 100 bp and it allows mismatches and gaps (Li and Durbin *et al* 2009). For mapping, the FASTQ files are needed as input. We used the default settings of BWA for mapping of paired-end reads.

The SAM (Sequence Alignment Map) format is the final output of BWA. Output and settings of the alignment tool are described in more detail in the material and methods section 3.8.2.1. For further analysis, SAMtools (Li *et al* 2009) was used to filter reads based on mapping quality and to remove PCR duplicates. In addition, SAMtools generates a pileup file which is a generic file format (3.8.2.2) and is used for downstream analyses.

BWA calculates the mapping quality scores for each alignment, which is the Phred-scaled probability of an alignment being incorrect. Phred quality values were originally developed by Phil Green and coworkers (Ewing *et al.* 1998). They integrated quality scores in a base calling program for automated sequencing traces. The principle of describing the quality of a base call was adopted for mapping qualities or SNP variants of next-generation sequencing data. The quality score Q is calculated from the probability P that the mapping or base is incorrect by the following formula: $Q = -10 \log_{10} P$. For example: a mapping score Q of 20 means that there is a 1 % chance that the mapping is incorrect. A Q score of 30 denotes a 0.1 % probability that the mapping is incorrect, or, with other words, the mapping is 99.9 % correct.

The Phred quality values are rounded to the nearest integer. In our workflow we removed reads with a mapping quality <10 using SAMtools. A mapping quality of 10 denotes an error probability of 0.1, so a mapping quality of 10 means that one out of 10 reads with that quality would be mapped to a wrong position. We also removed duplicate reads (SAMtools) introduced by PCR during library construction. SAMtools also detects reads with identical outer coordinates. In this case, only the read with the highest quality is retained. It is not possible to distinguish between duplicates resulting from PCR amplification or duplicates resulting from the same template by chance.

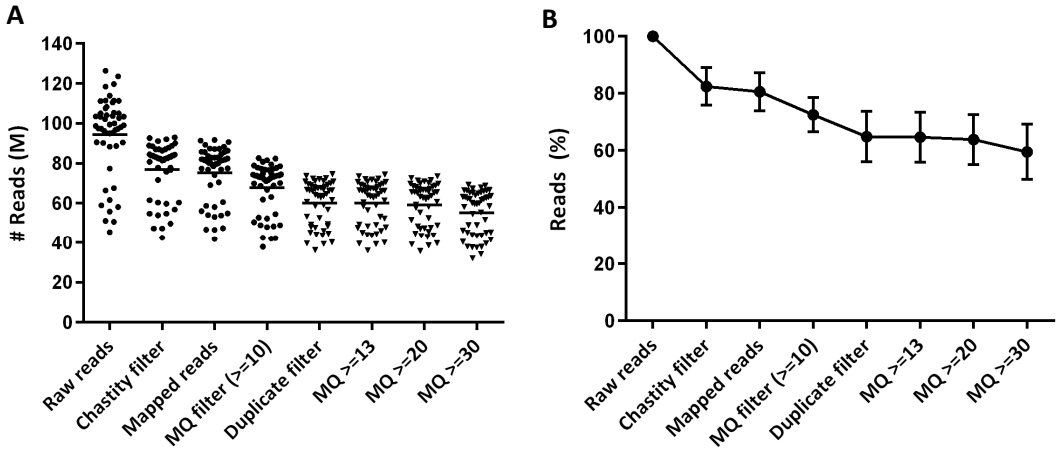


Figure 13 Filter settings and mapping quality for the reads from the 50 exomes. A) Total number of raw reads, reads filtered by chastity, mapped reads, mapping quality and reads after duplicate filtering. Unique reads mapped with a quality ≥10 were used for downstream analysis. Out of these the total number of reads with mapping quality ≥13, ≥20 and ≥30 is shown. ● Number of total reads to be further filtered. ▼ Number of total reads used for downstream analysis, separated by mapping quality. B) Mean (sd) percentage of reads mapped with the different filter settings, calculated for all 50 exomes. About 81.6 % of the reads passed the chastity filter. Further than that filter options for mapped reads and reads mapped with a quality of ≥10 were applied, resulting in 79.7 % and 71.8 % of the raw reads remaining, respectively. After removal of PCR duplicates 63.7 % of raw reads are remaining which were used for downstream analysis. 63.5 % of all reads mapped with a quality ≥13, 62.8 % mapped with quality ≥20 and 58.5 % had a mapping quality of 30 or better. MQ: mapping quality; M: million

On average 18.4 % of all reads were discarded by Illumina's chastity filter due to low signal to noise ratio (Figure 13). Reads passing this filter were mapped to the human genome reference by BWA. More than 97 % of all reads that passed the chastity filter could be mapped to a unique position in the genome. Mapping qualities were assigned to each read by the BWA program. A mapping quality of 0 indicates that a read can be mapped to more than one locus in the reference genome or is not mapped at all. These reads were discarded by the software. After mapping with BWA, the reads were further analyzed by SAMtools. The SAMtools's filter for mapping quality was set to ≥10 to retain reads mapped to the correct position with a probability of greater than 90 %. Between 8.9 to 10.8 % of the mapped reads did not fulfill this quality requirement. In addition, we removed on average 11 % duplicates from the reads after mapping quality filtering. Duplicate reads have exact the same 5´-end mapping coordinates. A duplicate read can either be the result of two PCR generated copies of the same genomic fragment being sequenced twice in different locations of a flow cell or it can be an optical duplicate when one cluster on the flow cell is erroneously identified as two adjacent clusters.

	Raw reads (M)	**Passed chastity filter** (M)	(%)	**Passed mapping quality (MQ≥10)** (M)	(%)	**Duplicate removal** (M)	(%)
Mean (sd)	94 (20)	77 (14)	82	68 (12)	72	60 (11)	64
Median (range)	99 (45-126)	83 (42-93)	84	73 (38-83)	74	64 (36-75)	65

Table 9 Number of reads after each filtering step. The percentage of reads remaining after the filtering steps was calculated based on the raw reads obtained. The minimum mapping quality was set at 10. The numbers shown are the mean and medians of all 50 exomes sequenced. MQ: mapping quality; M: million; sd: standard deviation

The mean (including the standard deviation: sd) and median (including the range) number of reads remaining after the different filtering steps is summarized in Table 9. The number of bases sequenced and used for downstream analysis was about 4.8 Gb per exome assuming a mean read length of 80 bp and considering all bases regardless of base quality.

For the reads, which were selected for downstream analysis, the distribution of mapping qualities was evaluated. The mapping quality tells us the probability that a read is mapped correctly. Table 10 shows the distribution of mapping qualities (after duplicate removal) for the 50 exomes sequenced in this project.

Mapping quality (# reads passed filter)	**≥MQ10** (M)	**≥MQ13** (M)	**≥MQ20** (M)	**≥MQ30** (M)
Mean (sd)	60.2 (11.2)	60.0 (11.1)	59.2 (11.0)	55.2 (11.2)
Median (range)	64.5 (36.1-74.6)	64.3 (36.0-74.4)	63.5 (35.6-73.6)	59.9 (32.0-69.2)
Mapping quality (Percentage of reads)	**≥MQ10 set to 100 %**	**≥MQ13** (%)	**≥MQ20** (%)	**≥MQ30** (%)
Mean (sd)	-	99.7 (0.0)	98.5 (0.2)	91.5 (4.3)

Table 10 Reads that passed the various filter criteria for mapping quality. The average number of reads per exome, which passed the filter with a certain mapping quality, is shown in the upper part of the table. In the lower part, the average percentage of reads (per exome) with a mapping quality ≥ 13, 20 and 30 was calculated from all the reads that reached a mapping quality of ≥ Q10. A total of 50 exomes were sequenced. M: million; MQ= Mapping Quality; sd: standard deviation

On average, 60.2 million reads per exome were available for downstream analysis. These reads remained after chastity and mapping quality filtering and after duplicate removal. To evaluate the mapping quality, we divided the mapped reads into those with a mapping quality ≥13, 20 and 30, which correspond to reads mapped with an error probability of less than 5, 1 and 0.1 %, respectively. On average, 60.0 million reads per exome had a mapping quality of 13 or higher and 59.2 million had a mapping quality of 20 or higher. More than 90 % of reads had a mapping quality of 30 or higher. A mapping quality of 30 indicates that there is a 1 in 1000 chance that the read is mapped to the wrong location.

4.1.2.2 *Enrichment Efficiency of the Agilent's SureSelect Kit*

Agilent's SureSelect kit consists of 120-mer biotinylated RNA baits for capturing genomic DNA fragments corresponding to exons. The proportions of reads or bases that map to target regions are a measure for the enrichment efficiency of the capture kit. We used the Integrative Genome Viewer (IGV) (Robinson *et al.* 2011; Thorvaldsdottir *et al.* 2012) to visualize the read mapping on the reference sequence. BAM files and tracks in the Bed file format can be loaded into the viewer and displayed in different tracks. The mapping data from one sample (P19) as well as the target regions of the Agilent 50 Mb kit and the hg18 RefSeq genes are shown as a screenshot of the IGV window in Figure 14.

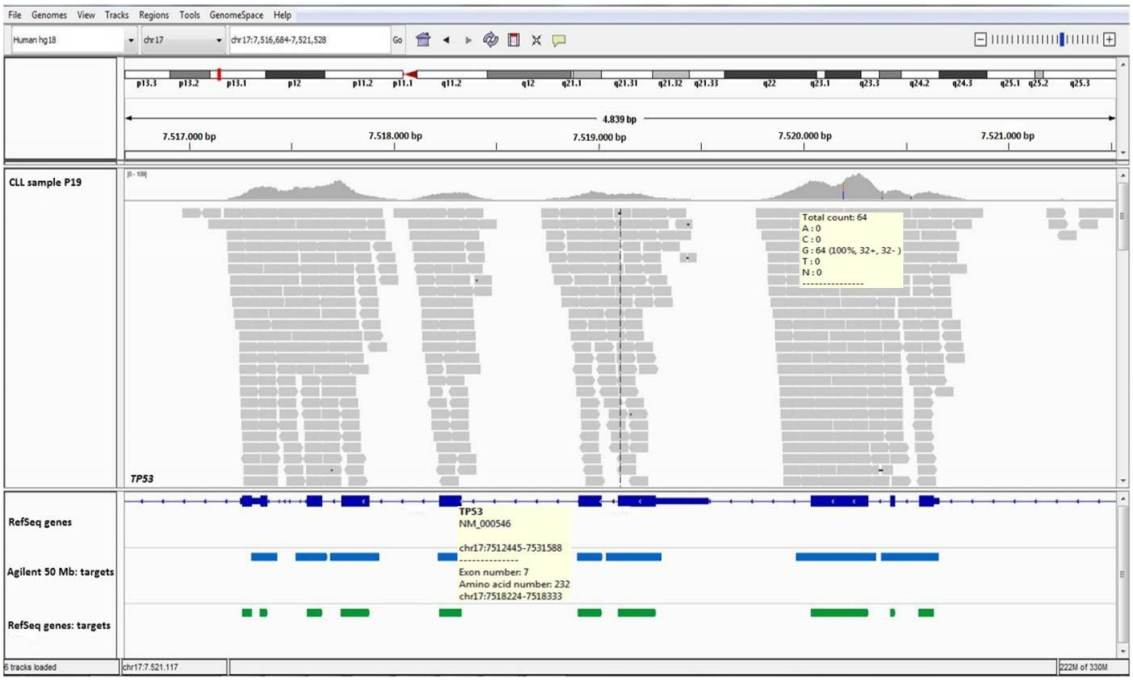


Figure 14 IGV screen shot. The data displayed in IGV shows the mapping of the reads from CLL sample P19 at the *TP53* locus. The coverage is displayed as a histogram in the upper portion of the "CLL sample P19" track and as individual reads in the lower portion of the track. Note that not all reads could be displayed because of lack of space. The annotation for the human genome is shown in dark blue and the Agilent SureSelect 50 Mb target regions in light blue. The RefSeq coding exons are displayed as green boxes in the bottom track.

The IGV program displays the mapping data in its genomic context and allows the user to zoom in and scroll over chromosomes. A chromosome ideogram is positioned at the top of the window, and the current position on the chromosome is indicated by a red box. In Figure 14, the mapped exome sequencing data of CLL sample P19 was loaded into the viewer, and the region around the *TP53* gene was magnified. The mapped reads are displayed as grey arrows indicating the sequencing direction of each individual read. The coverage is displayed as histogram in the upper portion of the window. The colored bars in the right histogram of the window indicate heterozygous SNPs (blue and orange). The human RefSeq gene track shown in dark blue with exons represented as boxes is displayed by default when the human genome is selected as reference. The Agilent 50 Mb targets (light blue) were uploaded. This track shows the position of the baits. The hg18 RefSeq coding exons are shown in the bottom track. By

moving the mouse pointer over the display additional information is shown in pale yellow text boxes.

The target positions of the Agilent 50 Mb kit (light blue boxes in Figure 14) overlap the exon boundaries of the RefSeq gene annotations. As a result of the position of the baits and the random fragments of DNA in the sequencing library, there are reads that map only partially or not at all to the targets. Some reads also map to intergenic regions due to sequence homologies or inefficient capturing. The performance of the Agilent 50 Mb kit for capturing the target regions was analyzed using a BED file provided by Agilent Technologies containing all target regions of the kit based on the NCBI Build 36 (hg18) annotation. However, for downstream analysis we used the human RefSeq genes as downloaded in May 2012 from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables). The RefSeq genes are based on the human genome assembly from March 2006 (NCBI Build 36/hg18). Only the RefSeq genes (34,380,348 bases) were considered to be the target in the downstream analysis if not otherwise specified.

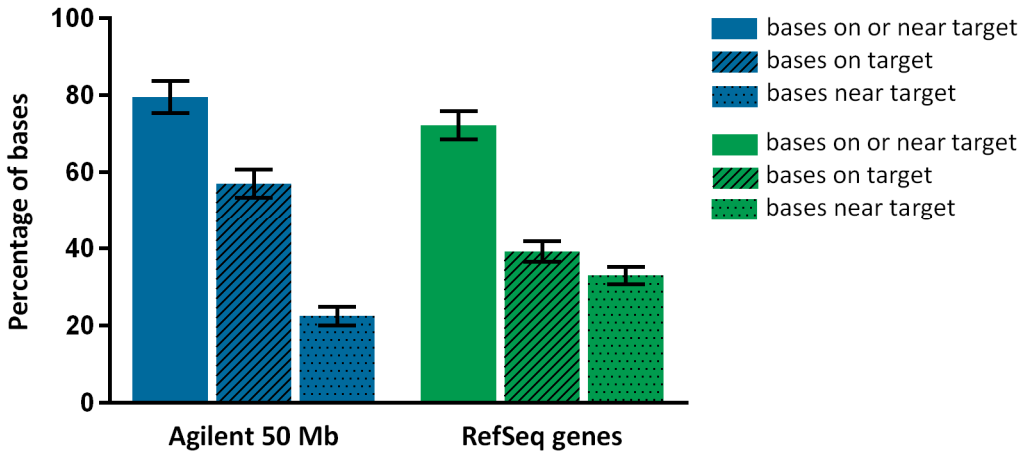


Figure 15 Quality of exome enrichment. On target metrics calculated at the single nucleotide level for the Agilent 50 Mb targets (blue). On average, 79.56 % of the bases map to or near to the target (± 250 nucleotides), 57.05 % mapped on the target and 22.51 % near to the target. When the RefSeq genes were considered as targets, fewer bases map on or near the target (72.72 %), on average 39.25 % map on target and 33.01 % close to the target (green). On average target metrics (including standard deviations) were calculated for all 50 exomes.

We used the Picard program (http://picard.sourceforge.net) to calculate on target metrics at the base level. We calculated the percentage of bases from mapped reads that mapped to target regions (on target) or within 250 nucleotides off target (near target). The percentage of on or near target bases when the RefSeq genes were used as target is with 72.72 % only a few percentage points lower than the 79.56 % when we used the Agilent 50 Mb kit as target. This difference is due to the fact that the Agilent 50 Mb bait library contains additional targets like miRNAs and transcripts annotated in other databases (CCDS, Ensembl and GenBank).

4.1.2.3 *Mean and Median on Target Coverage*

Coverage is defined as how many times a given genome position was sequenced. We calculated the coverage for the bases of the hg18 RefSeq genes. The coverage was calculated from a file generated by BEDtools, which contained the coverage at each nucleotide position of the target.

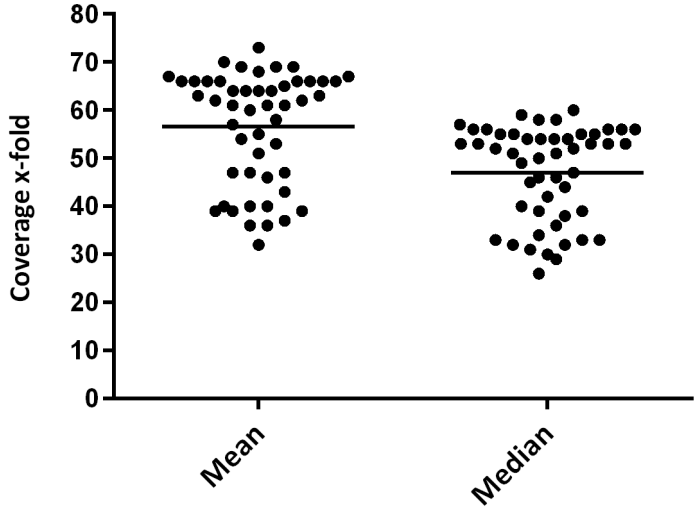


Figure 16 Mean and median on target coverage calculated for all 50 exomes. Each dot represents the mean or median coverage of one exome. The black line represents the mean.

For all 50 exomes, the average target coverage ranged between 32x and 73x with a mean of 57x (±11.7). The median coverage ranged from 26x to 60x with a mean of 47x (±9.8). The median coverage represents the data more accurate, as single positions with extraordinarily high coverage will not influence the median to the same extent as the mean.

4.1.2.4 *Percentage of Target Positions Sequenced*

It is of great interest to know how complete the exome target was sequenced. About 95 % of all target positions had a coverage of at least one. However, to be able to identify mutations or variants with a certain degree of confidence, a 10-fold coverage is required. A 10-fold coverage was achieved for about 85 % of all target positions in all 50 exomes. Half of all target positions had a coverage of more than 40x (Figure 17).

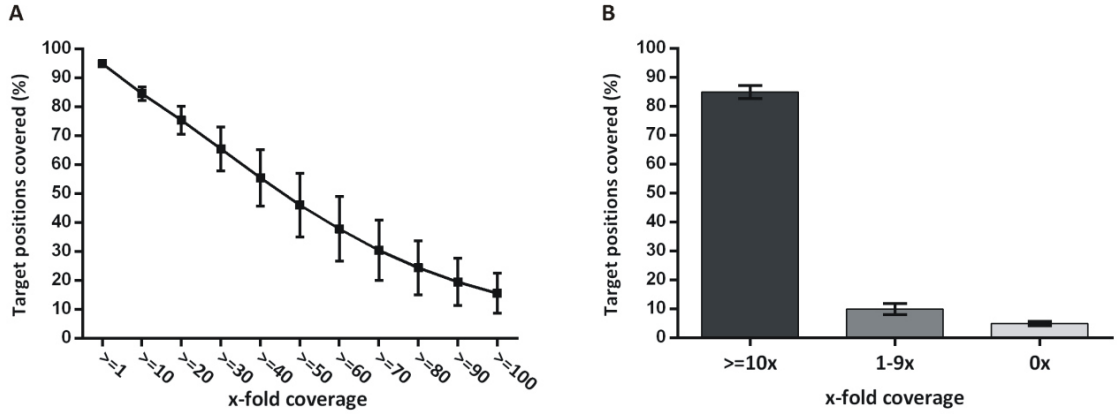


Figure 17 Coverage of target positions (hg18 RefSeq genes). A) Target positions genes covered x-fold or higher. B) Target positions covered ≥10x, 1-9x or not at all. The graph shows the coverage data for all 50 exomes, error bars represent standard deviation.

On average 5 % of all target regions were not covered at all in our 50 exome sequence samples. These targets were most likely not enriched during exome capture. On average, 10 % of all hg18 RefSeq positions were covered but were not covered by at least ten reads, which was our threshold for variant calling.

4.1.3 Downstream Analysis for the Detection of Single Nucleotide Variants (SNVs) and Insertion/Deletions (InDels)

One of the most important goals in the analysis of tumor genomes or exomes is the identification of tumor-specific genetic lesions. As a first step in this process, heterozygous positions in the sequence and positions where the sequence is different from a reference sequence have to be found.

Since 2009, a large number of variant detection programs have become available for the analysis of next-generation sequencing data. One of these is VarScan, which was developed by Dan Koboldt and colleagues (Koboldt *et al.* 2009). This tool is compatible with the output of short read aligners like BWA. There are two ways to use VarScan to detect tumor-specific variants: 1) in the so-called subtraction workflow, variants are first detected separately in the tumor and the remission sample, and then the remission variants are subtracted from the tumor variants: 2) in the so-called “somatic” workflow a built-in routine in VarScan 2, called somatic, is used, which directly identifies somatic mutations, loss of heterozygosity and copy number alterations in the sequencing data of paired tumor-normal samples (Koboldt *et al.* 2012). The results of the two workflows were compared to each other. From 2012 onward, we also used VarScan 2 for the subtraction workflow. VarScan 2 has an option to reduce false positive calls close to InDels. Many false positive calls occur adjacent to InDels as a result of faulty alignment because introducing a gap is much more “costly” than a mismatch.

4.1.3.1 *Subtraction Workflow to Detect CLL-specific Variants*

In the subtraction workflow, SNVs and InDels were called separately in the mapped read data for the CLL and the corresponding remission sample from the same patient. We used much more stringent filter settings for the variant calling in the CLL sample than in the remission sample. For the CLL sample a minimum coverage of 10 at the position of a putative variant was required. The minimum variant frequency was set to 20 %, and at least 3 variant reads had to be present. Since we attempted to validate all variants by Sanger sequencing, we assumed that a minimum variant frequency of 20 % would be the lowest detectable level using Sanger sequencing. Using these settings also implied that variants with a 20 % frequency would not be called at positions with a coverage between 10- and 14 fold since at least 3 variant reads were required. In other words, variants at a position with 10-14 fold coverage need at least a variant frequency of 21 % for a variant to be called.

In contrast to the CLL samples, the remission samples were analyzed with extremely relaxed filter settings. The coverage and the number of variant reads were both set at 1. We used a minimum variant frequency of 1 % in the high stringency (HS) setting and 5 % variant frequency in the low stringency (LS) setting (Figure 18). After variant calling, the variant lists from the tumor and normal sample were compared and only the variants present in the tumor sample and not in the remission sample were retained. To reduce the number of putative tumor-specific variants even further, all variants that were at the position of a known polymorphism (dbSNP130) and variants that were not located in the RefSeq coding exons were also excluded. We also discarded variants that were only supported by sequences from one strand and variants with a p-value >0.05 (Fisher's exact test). Custom filter settings retained only variants with at least four reads.

In a first step to identify those variants that might be so-called driver mutations, we annotated the effect of the CLL specific variants on the protein level using the snpEff program (Cingolani *et al.* 2012) to identify those variants that would result in a missense or nonsense mutation (nonsynonymous variants). InDels and nonsynonymous tumor-specific variants were manually inspected using the IGV browser. This inspection of candidate variants helped to further reduce the number of false positives calls resulting, for example from misalignments adjacent to germline InDels.

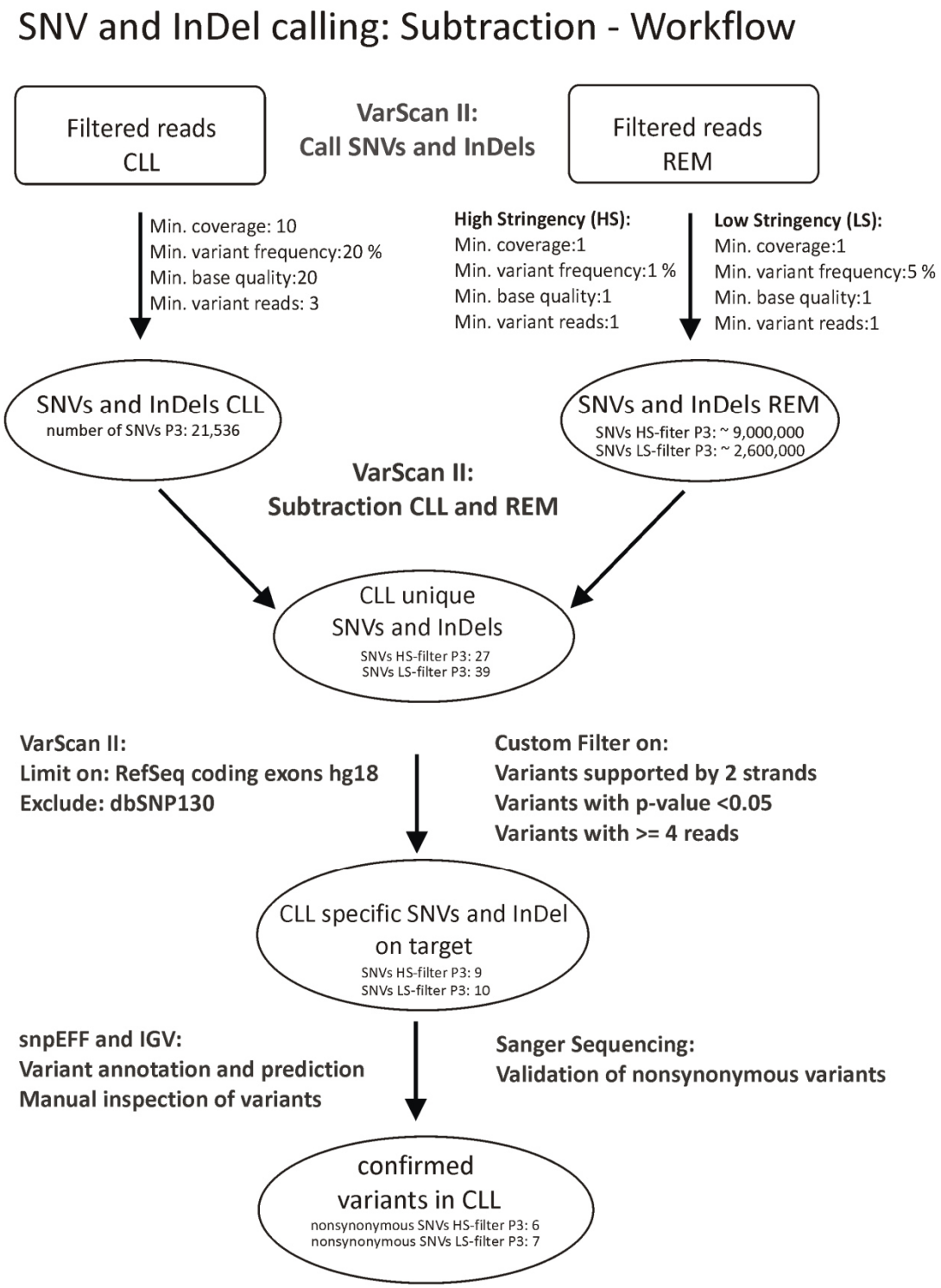


Figure 18 Flow chart 1 Subtraction workflow with high and low stringency filter settings using VarScan 2. Number of SNVs are given exemplary for P3. Among the nine and ten SNVs which remained after custom filter settings, six and seven were nonsynonymous for the HS and LS settings respectively. REM: remission sample; LS: Low Stringency filter settings; HS: High Stringency filter settings.

4.1.3.2 *Somatic Workflow to Detect CLL-specific Variants*

In the somatic workflow, using the VarScan 2 somatic subroutine, aligned sequencing data from a CLL and the corresponding remission sample were compared directly (Koboldt *et al.* 2012). At each position, the genotype is determined in the CLL and remission sample based on user-defined thresholds for base quality, coverage and variant frequency. The genotypes at each position are compared between tumor and normal sample. A variant is called homozygous when the reads supporting the variant are >75 % of the total reads at this position. If the variant genotypes of the CLL and the remission sample match at a given position, the variant is called germline (inherited). If the remission sample matches the reference and the tumor contains a variant, this position is called “somatic” indicating that this variant was probably somatically acquired. If the remission control is heterozygous at a given position and the tumor sample homozygous, this position is called “loss of heterozygosity” (LOH).

The following parameter settings were used for the somatic subroutine of VarScan 2: minimum coverage 10 for both tumor and remission sample minimum variant frequency for a heterozygous position to be called in the tumor sample was set to 20 %. Only variants supported by at least four reads were considered. A 20 % allele frequency and a minimum of four reads in the tumor sample means that variants at a position with 10-19 fold coverage need at least a variant frequency of 21 % for a variant to be called. High confidence variants were called with the following custom settings: (1) the maximum variant allele frequency in the remission sample was set to 5 %, (2) variants were required to have a p-value of <0.05 (Fisher’s exact test) (3) and reads from both strands of the tumor sample needed to be present at that position. Variant annotation and effect prediction were performed in the same way as in the subtraction workflow. In addition, all nonsynonymous variants were manually inspected by IGV before validation.

In contrast to the subtraction workflow, the somatic workflow also allows the user to enter the estimated purity of the tumor and normal sample. The LOH positions called by this workflow were manually inspected by IGV but not validated by Sanger sequencing. Due to time constraints these variants could not be analyzed in detail in this project.

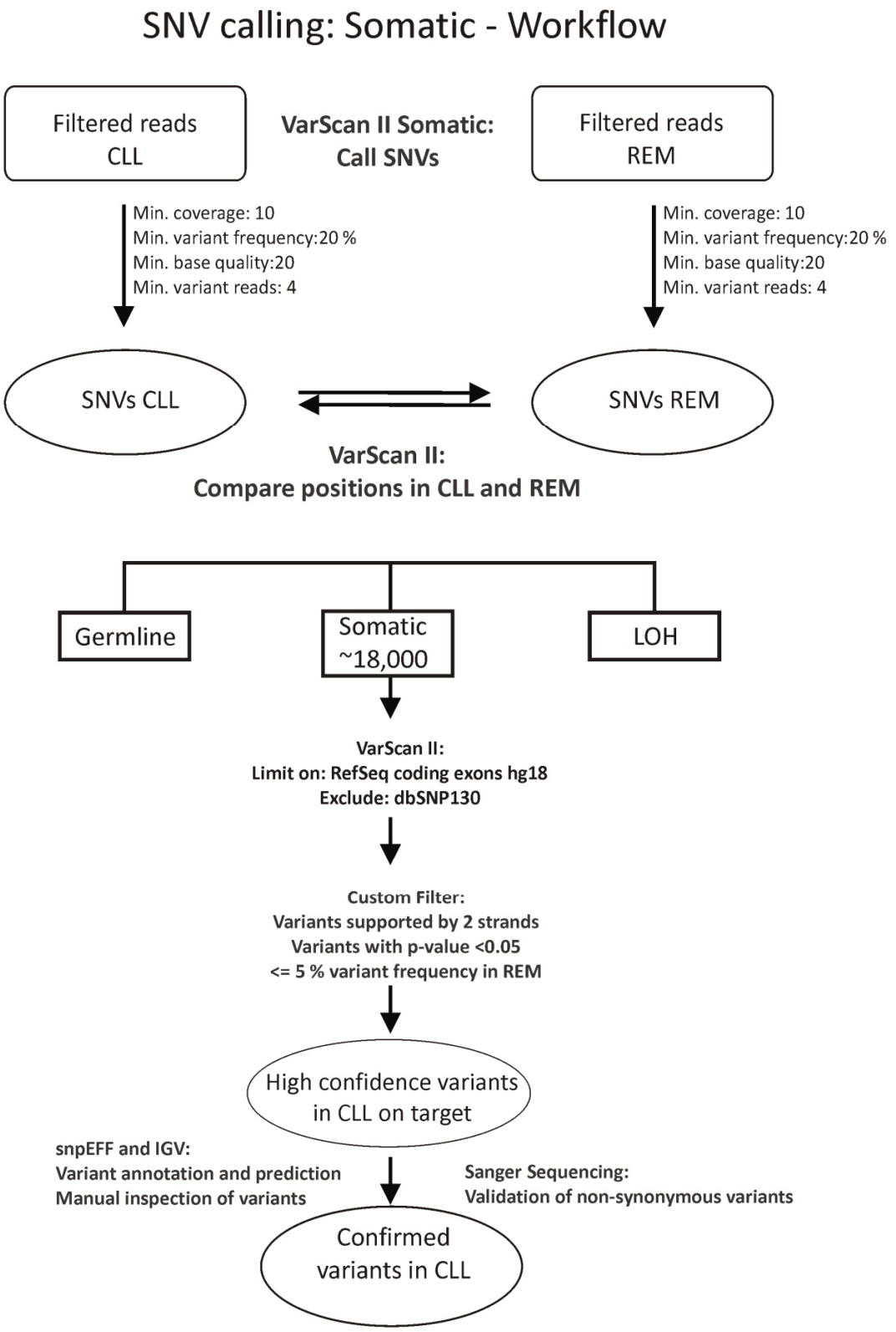


Figure 19 Somatic workflow using VarScan 2. REM: remission sample

4.1.3.3 *Validation by Sanger Sequencing*

To examine the rate of true positive calls, we performed Sanger sequencing (3.5.2) of all non-synonymous SNVs and InDels, which were called by our analysis pipeline. We set the minimum variant frequency to 20 % in our exome analysis pipeline, high enough to ensure that those candidates were detectable by Sanger sequencing if they were true positives. With a minimum variant frequency of 20 %, we also expect not to underestimate the number of true positive calls as we would not be calling possible mutations below the sensitivity of conventional sequencing. Besides data analysis, validation was one of the most laborious parts of the project. At the beginning we were uncertain which minimum requirements a true positive variant should fulfill. Therefore, we validated variant positions with different numbers of total reads, variant reads, various p-values and positions with variant reads from one or both strands. In total, we validated around 300 nonsynonymous SNVs and InDels. With this testing we became more familiar with the performance of the various filter settings and we then set the strict requirements for a variant to be validated as follows: p-value <0.05, ≥4 variant reads from both strands. For each amplicon we designed primers with a melting temperature of about 60 °C so that we could amplify multiple targets in one PCR run using a touch-down PCR protocol. The amplicon size ranged between 152 and 1065 bp for all variants that could be validated. To conserve genomic DNA from the original samples, we first tested the primers on DNA isolated from a cell line. In the next step we performed the PCR on the CLL as well as remission sample. In total, we performed around 600 PCR reactions on patient DNA. Starting from these amplicons, we further prepared twice the number of sequencing reactions for bidirectional Sanger sequencing.

4.1.3.4 *Comparing Subtraction Workflows with High and Low Stringency Settings*

To evaluate the differences between low and high stringency setting in the subtraction workflow we compared the number of CLL-specific SNVs that were detected in all 25 CLL exome samples with low *vs*. high stringency filter settings. The SNVs had to be supported by a minimum of four reads and a p-value <0.05. In addition, we only included nonsynonymous SNVs into this comparison, which were also resequenced by Sanger sequencing. We did not Sanger sequence candidate SNVs that did not result in a missense or nonsense mutation.

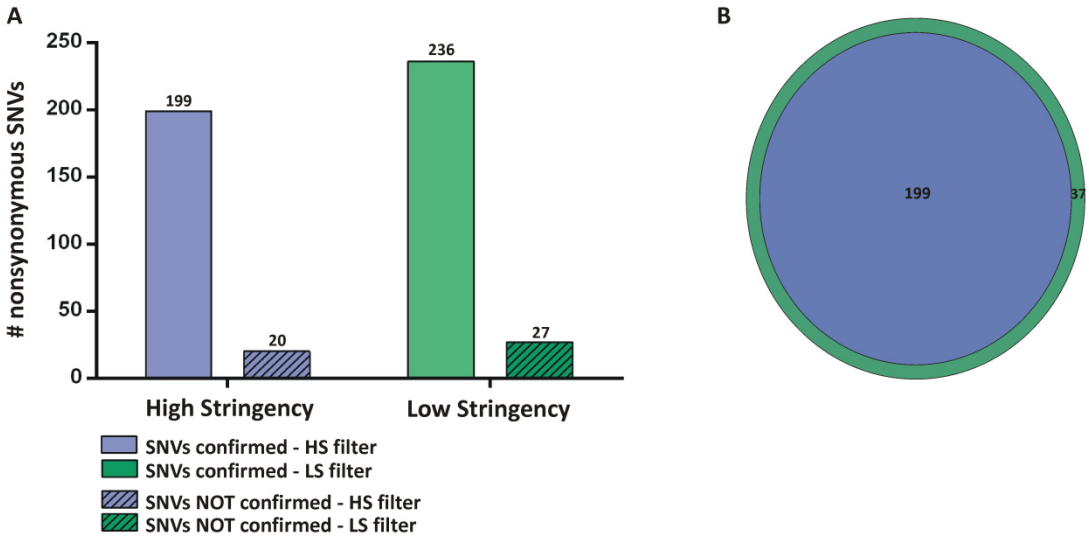


Figure 20 Comparing VarScan 2 subtraction workflows. A) In total n=199 validated mutations were detected using the high stringency filter and n=236 with low stringency filter settings. The number of SNVs that were false positives (non-confirmed) also increased from n=20 to n=27. B) All confirmed mutations detected with high stringency filter settings were also detected by the low stringency filter. HS: high stringency; LS: low stringency

The number of true positive nonsynonymous SNVs increased when we allowed the SNV to have a variant frequency of 5 % in the non-tumor sample (low stringency setting). This effectively means that we permitted cases, in which there were still some (less than about 10 %) CLL cells present in the remission sample. For the 25 CLL exomes the number of validated SNVs increased from n=199 to n=236 (Figure 20 A) when the low stringency settings were used. All SNVs detected by the high stringency (HS) filter settings were also detected by the low stringency (LS) filter (Figure 20 B). The number of variants that could not be confirmed by Sanger resequencing increased from n=20 to n=27 when the low stringency settings were applied (Figure 20 A). The SNVs, which could not be confirmed, were either false positives or SNVs which were also present in the remission sample. Some of the SNVs could already be excluded at the manual inspection stage with the IGV browser before Sanger sequencing. By increasing the variant frequency in the non-tumor sample to 5 %, the number of true positive of missense and nonsense mutations increases by 16 %, but the number of false positives increased by 35 %.

4.1.3.5 *Comparing the Subtraction Workflow with the Somatic Workflow*

We also compared the results of the somatic workflow and the low stringency subtraction workflow. For both workflows we only compared SNVs that could be confirmed by Sanger sequencing with a minimum of four variants in the tumor sample and a p-value of <0.05. The minimum variant frequency in tumor samples was set to 20 % in both workflows.

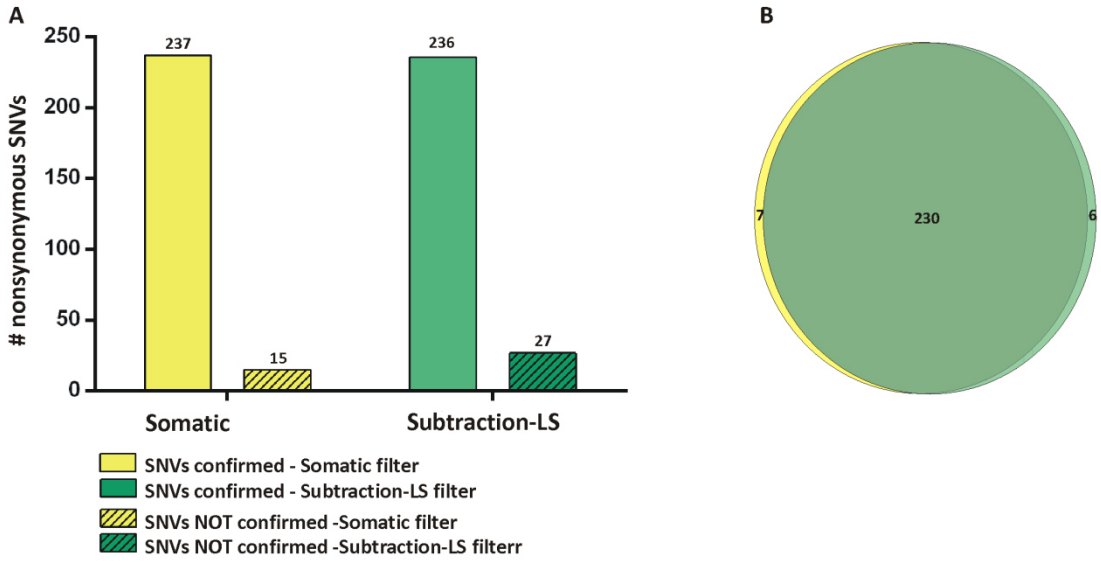


Figure 21 Comparison between the somatic and subtraction LS workflow. A) In total n=237 SNVs could be confirmed with the somatic filter settings, n=15 could not be confirmed. For the low stringency filter settings n=236 SNVs were confirmed, whereas n=27 proved to be false positives (not confirmed). B) Seven SNVs affecting seven genes could only be detected by the somatic workflow and six mutations in six genes were only identified with the low stringency subtraction workflow.

Almost equal numbers of true mutations were detected with the somatic and LS subtraction workflow, with n=237 and n=236 mutations, respectively (Figure 21). Although the amount of variants is similar, the variety of genes is higher. Seven genes were only observed within the somatic output and six were specific for the subtraction LS workflow (Figure 21). There were fewer false positive variants (n=15) with the somatic workflow compared to the LS subtraction workflow (n=27). It is important to note that three false positives SNVs in the LS subtraction workflow were called because of a low coverage (<10x) in the remission sample. This resulted from the minimum coverage setting of 1-fold in the LS subtraction workflow. The somatic workflow required a minimum coverage of 10 in the remission samples.

4.1.3.6 *Confirmed Missense and Nonsense Mutations in CLL*

Nonsynonymous SNVs resulting in a missense or nonsense mutation, which were obtained using the three different workflows, are summarized in Figure 22. All mutations shown were confirmed by Sanger sequencing of the CLL and remission sample (3.5.2). When we assign the nonsynonymous variants to the corresponding patient samples we see some variation in the number of mutations per sample. There were between 2 and 17 mutations per sample.

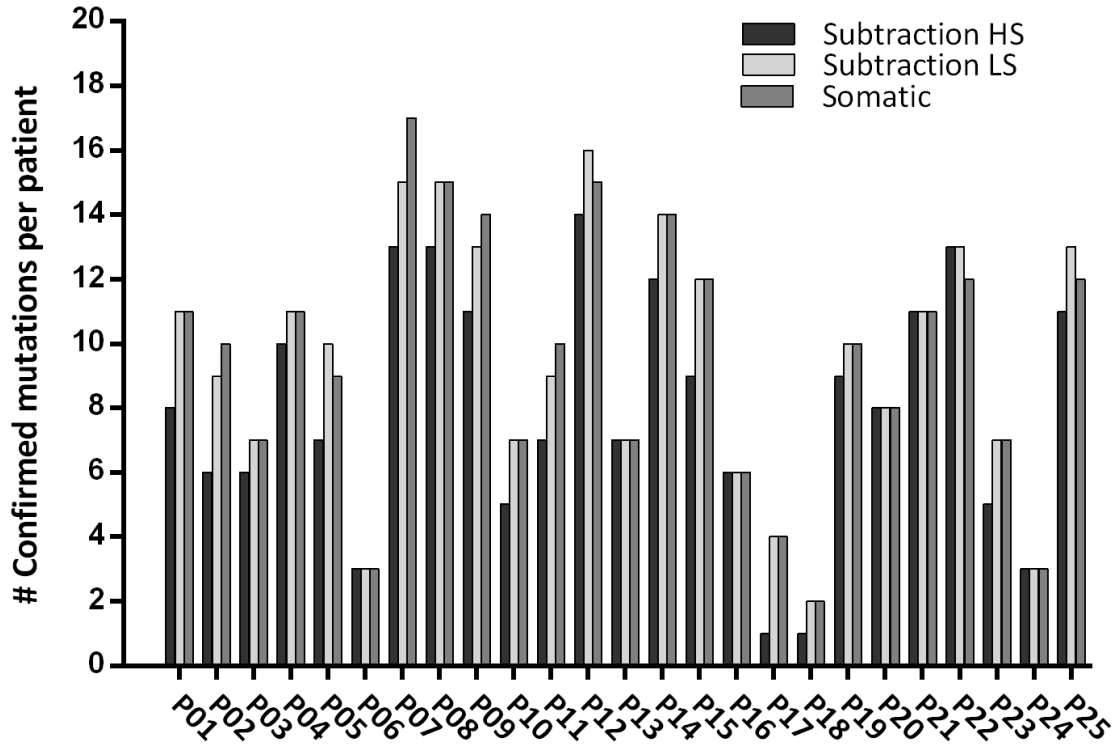


Figure 22 Missense and nonsense mutations detected in 25 CLL REACH exome samples. CLL patient samples from P01 – P25 and the number of their confirmed somatic mutations detected by the three workflows are plotted. HS: High stringency (minimum variant frequency in non-tumor sample 1 %); LS: Low stringency (minimum variant frequency in non-tumor sample 5 %)

Only mutations affecting the hg18 RefSeq genes (BED file) were analyzed and are shown here. Confirmed InDels are not included here. In all workflows, only SNVs with a p-value of <0.05 were considered for validations and the minimum variant frequency in the tumor sample was set to 20 %. The minimum coverage for a nucleotide position in a tumor sample was set to 10 in all workflows. In total, we were able to confirm 243 mutations with Sanger sequencing from all three workflows (199 in the Subtraction HS workflow; 236 in the Subtraction LS workflow; 237 in the Somatic workflow). The mutations detected with the low stringency settings included all mutations found with the high stringency setting (Figure 20 B). When we used the HS subtraction workflow we obtained between 1 and 14 mutations per patient. The mean was 8 (±3.8) mutations per patient. The LS subtraction workflow increased the mean number of mutations to 9 (±4.0) with a range of 2 and to 16 mutations. The somatic workflow produced similar results. The number of mutations ranged between 2 and 17, with a mean of 9 (±4.1). As mentioned earlier, about 15 % of all target positions could not be analyzed, as they did not reach the minimum coverage requirements for a SNV to be called. Therefore, the mutations described are derived from only 85 % of the target (i.e. the RefSeq hg18 genes). This implies that the true number of missense and nonsense mutations in each sample is about 17 % greater.

4.1.3.7 *Rate of True Positive Mutations in the Different Workflows*

Due to different filter settings and algorithms of the three workflows the rate of true positive missense and nonsense mutations varied between workflows. Of course, this "confirmation rate" is also dependent on the sensitivity of Sanger sequencing, which we used as the validation method. The proportion of true positives was calculated with respect to all non-synonymous variants detected with a given workflow. Some positions were excluded in advance, as they did not pass visual inspection using the IGV browser. The following graph represents the true positive rate of confirmed nonsynonymous variants for all three workflows from the 25 CLL samples (Figure 23).

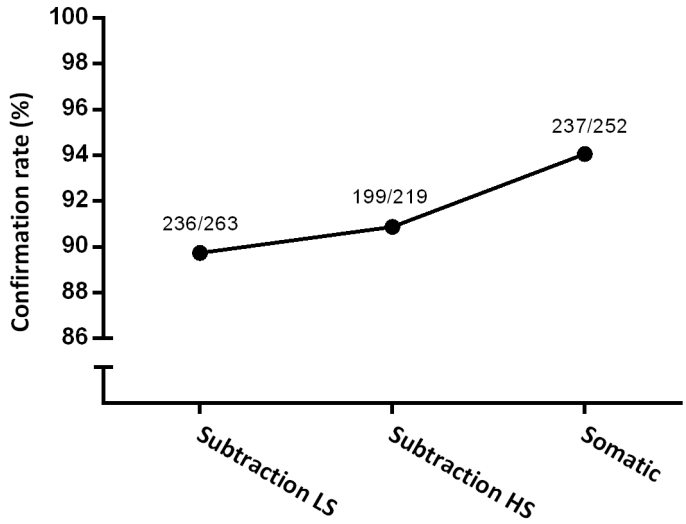


Figure 23 Frequency of variants confirmed by Sanger sequencing. The rate of true positive nonsynonymous calls was calculated from all nonsynonymous SNVs detected for each workflow after exclusion of low quality positions by visual inspection.

The highest confirmation rate (>94 %) was obtained with the somatic workflow followed by the high and low stringency subtraction workflows with a true positive rate of almost 91 % and 90 %, respectively (Figure 23). To calculate the confirmation rate for each workflow we only considered positions, which fulfilled strict criteria. Only variants supported by a minimum of four reads, a p-value of <0.05 and variants found on reads from both strands were considered. As our filter settings did not remove variants (but sort them), which did not fulfill these criteria, we also validated some of these SNV calls in the beginning of our project. Some of these could be confirmed but we did not consider them for this calculation, since we did not systematically validate SNVs, which did not fulfill these criteria in all samples. The validation rate for SNVs that did not pass these criteria was very low. The confirmation rate for InDels was also very low and was not calculated.

4.1.4 Somatic Mutations in CLL

In the 25 CLL samples that were whole exome sequenced, 271 mutations, which lead to an alteration at the protein level or affected a splice site region, were confirmed by Sanger sequencing. These mutations included mainly single nucleotide exchanges but also insertions

and deletions. These 271 mutations affected 260 different genes. The total number of mutations detected (n=271) is larger than reported in the preceding paragraphs because we also validated some putative SNVs by Sanger sequencing, which did not fulfill the strict requirements (p-value <0.05, ≥4 variant reads, both strands). These putative SNVs often had only three variant reads or reads from one strand only. The majority of these variants proved to be false positives. We mainly carried out this validation in the initial stages of the project to become familiar with the performance of the various filter settings in our workflows.

Missense	Nonsense	InDel	Read through	Splice site
237 (87.45 %)	19 (7.01 %)	13 (4.80 %)	1 (0.37 %)	1 (0.37 %)

Table 11 Mutations detected by exome sequencing and confirmed by Sanger sequencing. Total mutations n=271**.**

Among the validated variants, 87 % were missense mutations (n=237), 7 % nonsense mutations and 5 % insertions or deletions and 1 % read through and splice site mutations. Only one patient had a read through mutation in a stop codon. The splice site mutation was discovered by chance. We did not systematically look for splice site mutations.

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P01	GRHPR	9380	NM_012203.1	9	37420571	Missense	c.662C>T	p.A221V	40 %
P01	CHD2	1106	NM_001271.3	15	91335751	Missense	c.3455G>A	p.R1152Q	41 %
P01	FAR2	55711	NM_018099.4	12	29361161	Missense	c.1076C>A	p.A359D	41 %
P01	DMP1	1758	NM_004407.3	4	88799614	Missense	c.143G>A	p.S48N	56 %
P01	IL28A	282616	NM_172138.1	19	44451977	Missense	c.280C>T	p.R94C	46 %
P01	C1orf27	54953	NM_017847.5	1	184626488	STOP	c.497G>A	pW166*	54 %
P01	ARHGEF17	9828	NM_014786.3	11	72756343	Missense	c.6062G>A	p.R2021K	21 %
P01	FOLR4	390243	NM_001199206.1	11	93679364	Missense	c.176C>T	p.T59M	37 %
P01	PCDHGA4	56111	NM_018917.2	5	140715774	Missense	c.823G>A	p.G275R	46 %
P01	RUFY1	80230	NM_001040451.2	5	178961575	Missense	c.1514T>A	p.M505K	49 %
P01	ABCC10	89845	NM_033450.2	6	43525728	Missense	c.4316G>A	p.R1439H	67 %
P02	PAPD5	64282	NM_001040284.2	16	48807491	Missense	c.769A>T	p.I257F	38 %
P02	TCHHL1	126637	NM_001008536.1	1	150325863	Missense	c.919C>A	p.P307T	41 %
P02	CDH12	1010	NM_004061.3	5	22114366	Missense	c.177T>A	p.N59K	55 %
P02	ACOX2	8309	NM_003500.3	3	58492480	Missense	c.683A>G	p.Q228R	56 %
P02	ADAMTS12	81792	NM_030955.2	5	33624604	Missense	c.2722G>A	p.V908M	39 %
P02	CREBBP	1387	NM_004380.2	16	3869873	STOP	c.46A>T	p.K16*	73 %
P02	SRD5A1	6715	NM_001047.2	5	6716025	Missense	c.659C>G	p.A220G	49 %
P02	METRNL	284207	NM_001004431.1	17	78636203	Missense	c.271G>A	p.V91I	47 %
P02	OR4A47	403253	NM_001005512	11	48467438	Missense	c.518A>T	p.D173V	32 %
P02	ZNF227	7770	NM_182490.1	19	49432420	Missense	c.1997G>A	p.G666E	40 %
P02	RB1	5925	NC_000013.11	13	47845541	Splice Site	G>A	-	79 %
P03	SCUBE2	57758	NM_020974.2	11	9067848	Missense	c.238G>A	p.E80K	28 %
P03	MAGI2	9863	NM_012301.3	7	78920440	Missense	c.133G>T	p.G45W	37 %
P03	ATXN2L	11273	NM_148415.2	16	28754870	Missense	c.3011C>T	p.A1004V	46 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P03	RPS6KA3	6197	NM_004586.2	X	20121613	Missense	c.506A>G	p.D169G	29 %
P03	MED12	9968	NM_005120.2	X	70255979	Missense	c.131G>A	p.G44D	20 %
P03	C2orf67/ KANSL1L	151050	NM_152519.2	2	210727383	Missense	c.169A>C	p.T57P	26 %
P03	XPO1	7514	NM_003400.3	2	61572976	Missense	c.1711G>A	p.E571K	20 %
P03	EBF2	64641	NM_022659.3	8	25771883	Missense	c.1397C>T	p.T466M	46 %
P04	ADAMTS20	80070	NM_025003.3	12	42132649	Missense	c.1877A>T	p.N626I	54 %
P04	CYFIP2	26999	NM_014376.2	5	156718645	Missense	c.2728C>T	p.P910S	39 %
P04	XPO1	7514	NM_003400.3	2	61572975	Missense	c.1712A>G	p.E571G	53 %
P04	EMR2	30817	NM_013447.3	19	14744226	Missense	c.283G>A	p.V95M	37 %
P04	CSF2RB	1439	NM_000395.2	22	35664221	Missense	c.2425C>G	p.Q809E	48 %
P04	GRM6	2916	NM_000843.3	5	178341401	Missense	c.2497G>A	p.G833S	50 %
P04	RBPMS2	348093	NM_194272.1	15	62830842	Missense	c.136G>A	p.E46K	64 %
P04	SF3B1	23451	NM_012433.2	2	197975079	Missense	c.2098A>G	p.K700E	36 %
P04	CCDC111	201973	NM_152683.2	4	185840404	Missense	c.1016A>T	p.D339V	38 %
P04	CYB5D1	124637	NM_144607.4	17	7702519	Missense	c.234A>C	p.R78S	50 %
P04	PPP1R1B	84152	NM_032192.3	17	35045433	Missense	c.493C>T	p.R165C	34 %
P05	PCLO	27445	NM_033026.5	7	82422083	Missense	c.6122A>C	p.E2041A	40 %
P05	KAT5	10524	NM_006388.3	11	65240801	Missense	c.1018G>A	p.V340I	59 %
P05	CAPZA3	93661	NM_033328.2	12	18783343	Missense	c.874A>C	p.S292R	27 %
P05	IFT52	51098	NM_016004.2	20	41682917	Missense	c.706G>T	p.V236F	35 %
P05	MYH7	4625	NM_000257.2	14	22969669	Missense	c.939C>G	p.I313M	42 %
P05	TIAM2	26230	NM_012454.3	6	155492528	Missense	c.479G>A	p.R160Q	29 %
P05	BRDT	676	NM_207189.2	1	92219832	Missense	c.1934G>A	p.S645N	40 %
P05	ARHGEF5	7984	NM_005435.3	7	143706861	Missense	c.4605C>G	p.D1535E	46 %
P05	MYH2	4620	NM_001100112.1	17	10384070	Missense	c.1047G>T	p.K349N	32 %
P05	TRIML1	339976	NM_178556.3	4	189305147	Missense	c.1034T>A	p.V345E	40 %
P06	KIAA18437	285175	NM_032504.1	2	210495332	Missense	c.5444G>A	p.R1815H	44 %
P06	KLRC4	8302	NM_013431.2	12	10452793	Missense	c.271A>G	p.I91V	23 %
P06	GRK7	131890	NM_139209.2	3	142982234	Missense	c.941A>G	p.Y314C	29 %
P07	ANKS1B	56899	NM_001204081.1	12	97747087	Missense	c.80C>A	p.A27D	55 %
P07	FRAS1	80144	NM_025074.6	4	79590478	Missense	c.6424G>T	p.G2142C	44 %
P07	ROBO1	6091	NM_002941.3	3	78799773	Missense	c.1916C>T	p.A639V	50 %
P07	KIAA2022	340533	NM_001008537.2	X	73877566	Missense	c.3551G>C	p.S1184T	54 %
P07	KLHL6	89857	NM_130446.2	3	184755925	Missense	c.211C>G	p.L71V	59 %
P07	BAZ1A	11177	NM_013448.2	14	34313286	Missense	c.2995A>G	p.K999E	52 %
P07	MARCH6	10299	NM_005885.3	5	10458769	STOP	c.1432C>T	p.R478*	38 %
P07	UNC13C	440279	NM_001080534.1	15	52094479	Missense	c.2087A>T	p.E696V	55 %
P07	LRP1B	53353	NM_018557.2	2	141424319	Missense	c.3091T>G	p.C1031G	50 %
P07	OTUD7B	56957	NM_020205.2	1	148182861	Missense	c.2051G>T	p.R684M	40 %
P07	BMP2K	55589	NM_198892.1	4	80012800	Missense	c.1617G>T	p.Q539H	37 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P07	NBPF4	148545	NM_001143989.2	1	108573289	Missense	c.1436C>T	p.A479V	28 %
P07	LIMD1	8994	NM_014240.2	3	45611380	Missense	c.5A>G	p.D2G	38 %
P07	KLHL6	89857	NM_130446.2	3	184692708	Missense	c.1567G>T	p.G523W	36 %
P07	C12orf51/ HECTD4	283450	NM_001109662.3	12	111130155	STOP	c.7635G>A	p.W2545*	48 %
P07	COL18A1	80781	NM_030582.3	21	45731807	Missense	c.2392G>A	p.G798R	50 %
P07	NOX1	27035	NM_007052.4	X	99991879	Missense	c.1050C>A	p.F350L	58 %
P08	TTN	7273	NM_133378.4	2	179154538	Missense	c.58999 G>A	p.V19667I	49 %
P08	LRRC7	57554	NM_020794.2	1	70278036	Missense	c.3827G>A	p.R1276K	43 %
P08	TNFAIP3	7128	NM_006290.3	6	138234147	Missense	c.90T>A	p.F30L	48 %
P08	PASD1	139135	NM_173493.2	X	150593064	Missense	c.1925C>T	p.P642L	52 %
P08	DOCK10	55619	NM_014689.2	2	225366381	Missense	c.5192C>T	p.A1731V	30 %
P08	TNS3	64759	NM_022748.11	7	47406532	Missense	c.902C>T	p.T301M	57 %
P08	GRIA2	2891	NM_000826.3	4	158500511	Missense	c.2057C>T	p.A686V	44 %
P08	LRFN5	145581	NM_152447.3	14	41430859	Missense	c.2042G>A	p.R681H	28 %
P08	MYO6	4646	NM_004999.3	6	76607703	Missense	c.704T>A	p.I235N	22 %
P08	FREM2	341640	NM_207361.4	13	38331672	Missense	c.7464T>A	p.N2488K	31 %
P08	CELSR3	1951	NM_001407.2	3	48658567	Missense	c.7423C>T	p.R2475W	43 %
P08	RBM46	166863	NM_144979.4	4	155939514	Missense	c.750A>C	p.E250D	29 %
P08	DIO2	1734	NM_013989.4	14	79747441	Missense	c.128G>A	p.R43H	48 %
P08	ANKFN1	162282	NM_153228.2	17	51881489	Missense	c.1159G>T	p.V387F	25 %
P08	PDE1C	5137	NM_001191058.1	7	31885200	Missense	c.539G>A	p.R180K	53 %
P09	TOP2A	7153	NM_001067.3	17	35810749	Missense	c.2543T>C	p.V848A	47 %
P09	MRO	83876	NM_031939.3	18	46581740	Missense	c.562G>T	p.D188Y	42 %
P09	ASH1L	55870	NM_018489.2	1	153714418	Missense	c.4868G>A	p.G1623S	48 %
P09	ARSK	153642	NM_198150.2	5	94962507	Missense	c.1297G>A	p.A433T	41 %
P09	SKIV2L2	23517	NM_015360.4	5	54684790	Missense	c.1469A>G	p.N490S	32 %
P09	SIPA1L1	26037	NM_015556.1	14	71125565	Missense	c.1223A>T	p.N408I	47 %
P09	PTPRT	11122	NM_007050.5	20	40414194	Missense	c.1706G>A	p.S569N	28 %
P09	ADAM8	101	NM_001109.4	10	134934461	Missense	c.1478A>G	p.E493G	50 %
P09	SCD5	79966	NM_001037582.2	4	83845516	Missense	c.307C>T	p.R103W	28 %
P09	ADAMTS14	140766	NM_080722.3	10	72138406	Missense	c.736G>A	p.D246N	31 %
P09	AMHR2	269	NM_020547.2	12	52104426	Missense	c.137G>T	p.G46V	39 %
P09	SCRN2	90507	NM_138355.3	17	43270693	Missense	c.1061G>A	p.R354H	48 %
P09	DNAH5	1767	NM_001369.2	5	13894872	Missense	c.5413C>T	p.R1805C	35 %
P09	MYH2	4620	NM_001100112.1	17	10389453	Missense	c.440A>G	p.K147R	45 %
P10	RANBP2	5903	NM_006267.4	2	108766484	Missense	c.9370G>A	p.G3124R	38 %
P10	ERCC8	1161	NM_000082.3	5	60231264	Missense	c.665G>A	p.C222Y	43 %
P10	AKR1C4	1109	NM_001818.3	10	5232219	Missense	c.160T>A	p.L54I	29 %
P10	PRKCA	5578	NM_002737.2	17	62169255	Missense	c.1439C>T	p.A480V	25 %
P10	COL5A2	1290	NM_000393.3	2	189616186	Missense	c.3407G>A	p.R1136Q	39 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P10	KLRC2	3822	NM_002260.3	12	10477782	Missense	c.358G>A	p.E120K	75 %
P10	KIF9	64147	NM_001134878.1	3	47261974	Missense	c.1321C>T	p.R441C	27 %
P11	ZFP161/ ZBTB14	7541	NM_003409.4	18	5281893	Missense	c.314T>A	p.V105D	39 %
P11	ARMC4	55130	NM_018076.2	10	28265762	Missense	c.2151G>C	p.K717N	36 %
P11	MRE11A	4361	NM_005591.3	11	93836981	Missense	c.1171A>T	p.N391Y	23 %
P11	NRBP1	29959	NM_013392.2	2	27509813	STOP	c.169G>T	p.E57*	52 %
P11	KCNJ5	3762	NM_000890.3	11	128291514	Missense	c.938G>A	p.G313D	32 %
P11	FAM179A	165186	NM_199280.2	2	29103242	Missense	c.1873C>T	p.R625W	46 %
P11	ATM	472	NM_000051.3	11	107741297	Missense	c.9023G>A	p.R3008H	22 %
P11	GPR61	83873	NM_031936.4	1	109887409	Missense	c.242A>G	p.H81R	44 %
P11	EGR2	1959	NM_001136178.1	10	64243338	Missense	c.1066G>A	p.E356K	39 %
P11	SLC2A2	6514	NM_000340.1	3	172207696	Missense	c.547G>A	p.A183T	44 %
P12	KIAA0240	23506	NM_015349.1	6	42905325	Missense	c.1276C>G	p.Q426E	31 %
P12	RNASE2	6036	NM_002934.2	14	20494104	Missense	c.334C>T	p.L112F	43 %
P12	LGR5	8549	NM_003667.3	12	70233271	Missense	c.580G>T	p.A194S	42 %
P12	ABCA9	10350	NM_080283.3	17	64543452	Missense	c.886G>A	p.V296I	49 %
P12	KIAA1586	57691	NM_020931.2	6	57026109	Missense	c.853C>T	p.R285C	41 %
P12	KIAA1712/ CEP44	80817	NM_001145314.1	4	175462073	Missense	c.485G>A	p.G162D	36 %
P12	CDH12	1010	NM_004061.3	5	21890556	Missense	c.627C>A	p.F209L	51 %
P12	ADAM17	6868	NM_003183.4	2	9562809	Missense	c.1481T>G	p.M494R	44 %
P12	PAM	5066	NM_000919.3	5	102338007	Missense	c.1451G>A	p.G484D	36 %
P12	RIMS2	9699	NM_001100117.2	8	105095968	STOP	c.3283C>T	p.R1095*	28 %
P12	PTPRU	10076	NM_133178.3	1	29457762	Missense	c.364C>T	p.R122C	48 %
P12	FLNB	2317	NM_001457.3	3	58084289	Missense	c.3556A>C	p.K1186Q	41 %
P12	CASP6	839	NM_001226.3	4	110831463	Missense	c.635T>A	p.V212D	37 %
P12	PACS2	23241	NM_015197.3	14	104904619	Missense	c.448G>A	p.G150S	46 %
P12	CTAGE5	4253	NM_005930.3	14	38833003	Missense	c.544T>C	p.S182P	40 %
P12	PRPS1L1	221823	NM_175886.2	7	18033893	Missense	c.38A>G	p.Q13R	20 %
P12	ACSM2B	348158	NM_182617.3	16	20465291	Missense	c.1105A>T	p.T369S	48 %
P13	OR5W2	390148	NM_001001960.1	11	55437838	Missense	c.797A>G	p.Y266C	57 %
P13	BHLHB9	80823	NM_001142525.1	X	101891757	Missense	c.1178C>A	p.S393Y	49 %
P13	GANC	2595	NM_198141.2	15	40419269	STOP	c.1954C>T	p.R652*	43 %
P13	GTSF1	121355	NM_144594.2	12	53145216	Missense	c.19G>A	p.D7N	39 %
P13	HPSE	10855	NM_001098540.2	4	84453395	Missense	c.569C>T	p.A190V	39 %
P13	C20orf54/ SLC52A3	113278	NM_033409.3	20	694403	Missense	c.16C>T	p.H6Y	36 %
P13	IGFN1	91156	NM_001164586.1	1	199461632	Missense	c.10544C>T	p.T3515M	30 %
P13	MYL12A	10627	NM_006471.2	18	3243964	Missense	c.259A>G	p.T87A	53 %
P13	FAM32A	26017	NM_014077.2	19	16162758	Missense	c.331A>G	p.T111A	39 %
P14	PLS1	5357	NM_001145319.1	3	143872568	Missense	c.278G>A	p.R93Q	44 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P14	OR2L8	391190	NM_001001963.1	1	246179686	STOP	c.904C>T	p.R302*	42 %
P14	SLC6A5	9152	NM_004211.3	11	20616662	Missense	c.1951G>A	p.G651R	48 %
P14	OR5J2	282775	NM_001005492.1	11	55701154	Missense	c.485G>A	p.S162N	39 %
P14	SLC6A11	6538	NM_014229.1	3	10949837	Missense	c.1372G>T	p.G458C	39 %
P14	FLNC	2318	NM_001458.4	7	128284512	Missense	c.7666C>T	p.P2556S	46 %
P14	FUBP1	8880	NM_003902.3	1	78217236	STOP	c.41C>G	p.S14*	37 %
P14	CCDC85A	114800	NM_001080433.1	2	56273648	Missense	c.809G>A	p.R270H	38 %
P14	ZNF512B	57473	NM_020713.1	20	62061926	Missense	c.2438G>T	p.R813L	54 %
P14	ADAMTS3	9508	NM_014243.2	4	73373443	Missense	c.2938G>A	p.V980M	50 %
P14	ARSD	414	NM_001669.3	X	2835439	Missense	c.1655G>A	p.R552Q	50 %
P14	CCDC56	28958	NM_001040431.2	17	38204198	Missense	c.28C>A	p.L10M	26 %
P14	COL4A3	1285	NM_000091.4	2	227850418	Missense	c.2030G>A	p.G677E	22 %
P14	GRIK1	2897	NM_000830.3	21	29988086	Missense	c.286G>A	p.A96T	40 %
P14	SMCHD1	23347	NM_015295.2	18	2678396	Missense	c.643C>A	p.H215N	46 %
P14	XPO1	7514	NM_003400.3	2	61572976	Missense	c.1711G>A	p.E571K	7 %
P15	DNAH7	56171	NM_018897.2	2	196428797	Missense	c.8578G>T	p.A2860S	47 %
P15	SLIT2	9353	NM_004787.1	4	19867412	Missense	c.199A>T	p.I67F	54 %
P15	XPR1	9213	NM_004736.3	1	179061080	Missense	c.1111C>T	p.R371W	40 %
P15	KRT20	54474	NM_019010.2	17	36292425	Missense	c.398A>G	p.D133G	53 %
P15	L2HGDH	79944	NM_024884.2	14	49838577	Missense	c.316G>A	p.E106K	41 %
P15	MYEF2	50804	NM_016132.3	15	46237482	Missense	c.985C>T	p.R329C	45 %
P15	RBM25	58517	NM_021239.2	14	72613953	Missense	c.280A>T	p.I94F	49 %
P15	KRT4	3851	NM_002272.3	12	51487873	Missense	c.1168G>A	p.G390S	42 %
P15	XPO1	7514	NM_003400.3	2	61572976	Missense	c.1711G>A	p.E571K	30 %
P15	FLJ32682/ FAM194B	220081	NM_182542.2	13	45059211	Missense	c.844G>A	p.D282N	67 %
P15	SF3B1	23451	NM_012433.2	2	197975066	Missense	c.2111T>A	p.I704N	42 %
P15	ITIH4	3700	NM_002218.4	3	52832920	Missense	c.1312C>T	p.R438W	65 %
P15	PCLO	27445	NM_033026.5	7	82419594	Missense	c.8611C>T	p.P2871S	43 %
P16	DST	667	NM_015548.4	6	56528572	Missense	c.6797G>A	p.R2266Q	41 %
P16	ARHGAP18	93663	NM_033515.2	6	129996915	Missense	c.581T>G	p.L194R	42 %
P16	DEFB114	245928	NM_001037499.1	6	50036079	Missense	c.95G>A	p.R32H	47 %
P16	DOCK1	1793	NM_001380.3	10	129121682	Missense	c.4997G>T	p.G1666V	26 %
P16	OBSCN	84033	NM_001271223.2	1	226588071	Missense	c.18892G>T	p.A6298S	39 %
P16	ASMTL	8623	NM_001173474.1	X/Y	1482164	Extension	c.1816T>G	*606G	100 %
P16	LAMA5	3911	NM_005560.3	20	60327354	Missense	c.6982G>A	p.A2328T	30 %
P16	AKR1B1	231	NM_001628.2	7	133785062	STOP	c.379G>T	p.E127*	42 %
P17	DNAJC13	23317	NM_015268.3	3	133661884	Missense	c.1550C>A	p.S517Y	39 %
P17	ICAM5	7087	NM_003259.3	19	10263792	Missense	c.755A>G	p.D252G	36 %
P17	LAMA1	284217	NM_005559.3	18	6955337	Missense	c.7145A>G	p.Y2382C	40 %
P17	GRIN2A	2903	NM_001134407.1	16	9939789	Missense	c.535G>A	p.G179S	33 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P17	SLC26A5	375611	NM_198999.2	7	102838136	Missense	c.667A>G	p.T223A	46 %
P18	GPC6	10082	NM_005708.3	13	93832730	Missense	c.1214C>A	p.T405N	22 %
P18	DUSP13	51207	NM_001007272.1	10	76527551	Missense	c.248G>A	p.R83H	27 %
P18	PDE2A	5138	NM_002599.4	11	71977494	Missense	c.1052A>G	p.N351S	22 %
P18	SIAH3	283514	NM_198849.2	13	45255947	Missense	c.382C>T	p.R128W	25 %
P19	APOB	338	NM_000384.2	2	21084498	Missense	c.8747C>T	p.A2916V	57 %
P19	APOB	338	NM_000384.2	2	21084417	Missense	c.8828A>G	p.E2943G	57 %
P19	MYO18B	84700	NM_032608.5	22	24753166	Missense	c.7226G>A	p.R2409H	42 %
P19	STX17	55014	NM_017919.2	9	101753188	Missense	c.215T>C	p.I72T	42 %
P19	FBXO47	494188	NM_001008777.2	17	34352567	Missense	c.1073T>C	p.F358S	57 %
P19	FGF14	2259	NM_175929.2	13	101173307	STOP	c.634C>T	p.R212*	48 %
P19	ADAMTS13	11093	NM_139025.3	9	135279277	Missense	c.188C>T	p.P63L	41 %
P19	IL3RA	3563	NM_002183.3	X	1431348	Missense	c.565G>A	p.A189T	41 %
P19	KRAS	3845	NM_004985.3	12	25289551	Missense	c.35G>A	p.G12D	42 %
P19	CHPF	79586	NM_024536.5	2	220113433	Missense	c.1244G>A	p.R415H	29 %
P19	TKTL2	84076	NM_032136.4	4	164613013	STOP	c.1324C>T	p.R442*	51 %
P20	SF3B1	23451	NM_012433.2	2	197975618	Missense	c.1984C>G	p.H662D	48 %
P20	HOOK1	51361	NM_015888.4	1	60102919	Missense	c.1654G>A	p.A552T	41 %
P20	OR9Q1	219956	NM_001005212.3	11	57703997	Missense	c.505T>C	p.C169R	38 %
P20	PCSK1	5122	NM_000439.4	5	95787327	Missense	c.349G>A	p.A117T	38 %
P20	CDK20	23552	NM_001170640.1	9	89778764	Missense	c.82G>A	p.E28K	37 %
P20	USP8	9101	NM_001128610.1	15	48569368	Missense	c.1969A>G	p.K657E	37 %
P20	SF4/SUGP1	57794	NM_172231.3	19	19282002	Missense	c.214A>G	p.N72D	50 %
P20	C3orf15/ MAATS1	89876	NM_033364.3	3	120945710	Missense	c.1879G>A	p.E627K	39 %
P20	SCEL	8796	NM_003843.3	13	77082191	Missense	c.1226G>A	p.S409N	35 %
P21	VEGFC	7424	NM_005429.2	4	177845450	Missense	c.1030A>C	p.K344Q	37 %
P21	MLH3	27030	NM_001040108.1	14	74583229	Missense	c.2883G>T	p.E961D	37 %
P21	TESK2	10420	NM_007170.2	1	45695996	STOP	c.49G>T	p.E17*	35 %
P21	EYS	346007	NM_001142800.1	6	64488697	Missense	c.9189T>A	p.S3063R	21 %
P21	DSP	1832	NM_001008844.1	6	7524576	Missense	c.3154T>A	p.C1052S	30 %
P21	ZDHHC20	253832	NM_153251.3	13	20853671	Missense	c.959C>G	p.P320R	27 %
P21	SCN10A	6336	NM_006514.2	3	38738783	Missense	c.3477C>G	p.I1159M	29 %
P21	TRAF3	7187	NM_145726.2	14	102441342	Missense	c.1100T>C	p.L392P	33 %
P21	TP53	7157	NM_000546.5	17	7518996	Missense	c.578A>G	p.H193R	38 %
P21	FASTK	10922	NM_033015.3	7	150406598	Missense	c.386C>T	p.T129M	20 %
P21	CXCR4	7852	NM_003467.2	2	136588955	STOP	c.1013C>G	p.Ser338*	20 %
P21	HEATR3	55027	NM_182922.2	16	48691642	Missense	c.1600T>C	p.C534R	21 %
P21	CHRD	8646	NM_003741.2	3	185587355	Missense	c.2225T>C	p.V742A	31 %
P22	DOCK4	9732	NM_014705.3	7	111272105	STOP	c.2686C>T	p.R896*	54 %
P22	CSMD1	64478	NM_033225.5	8	3253055	Missense	c.1844T>C	p.I615T	48 %

Case	Gene Name	Gene ID	Annotation	Chr	Position	Type of Mutation	cDNA change	Protein change	Var. Freq.
P22	DDX3X	1654	NM_001356.3	X	41090557	Missense	c.1447G>A	p.A483T	97 %
P22	COL11A1	1301	NM_001854.3	1	103261064	Missense	c.1067A>T	p.E356V	47 %
P22	TSHZ2	128553	NM_001193421.1	20	51304068	Missense	c.655G>A	p.A219T	69 %
P22	FAT3	120114	NM_001008781.2	11	92255907	Missense	c.12637C>T	p.R4213C	40 %
P22	AHSG	197	NM_001622.2	3	187821328	Missense	c.1019G>A	p.R340H	46 %
P22	KMO	8564	NM_003679.4	1	239794923	Missense	c.629C>G	P210R	33 %
P22	KRAS	3845	NM_004985.3	12	25289548	Missense	c.38G>A	pG13D	52 %
P22	UCHL1	7345	NM_004181.4	4	40960052	Missense	c.513C>G	p.H171Q	26 %
P22	WDR33	55339	NM_018383.4	2	128182867	Missense	c.3635C>G	p.A1212G	22 %
P22	SETD2	29072	NM_014159.6	3	47114571	STOP	c.5020G>T	p.E1674*	58 %
P23	NRAS	4893	NM_002524.4	1	115060267	Missense	c.38G>T	p.G13V	29 %
P23	CACNB1	782	NM_000723.4	17	34585240	Missense	c.1529C>G	p.S510C	35 %
P23	MCC	4163	NM_002387.2	5	112448828	Missense	c.907C>T	p.R303C	31 %
P23	PRMT1	3276	NM_198318.4	19	54877015	Missense	c.175C>T	p.R59C	27 %
P23	TBCC	6903	NM_003192.2	6	42821000	Missense	c.790C>T	p.H264Y	20 %
P23	HERPUD2	64224	NM_022373.4	7	35640449	Missense	c.1057A>G	p.M353V	29 %
P23	DDX43	55510	NM_018665.2	6	74172247	Missense	c.775A>C	p.K259Q	23 %
P23	UMOD	7369	NM_003361.2	16	20267895	Missense	c.229T>A	p.C77S	40 %
P24	MURC	347273	NM_001018116.1	9	102380607	STOP	c.361C>T	p.Q121*	48 %
P24	ZNF292	23036	NM_015021.1	6	88023170	Missense	c.3104A>G	p.N1035S	29 %
P24	MSH6	2956	NM_000179.2	2	47886308	Missense	c.3604A>T	p.M1202L	34 %
P25	FAT4	79633	NM_024582.4	4	126631124	Missense	c.13697G>A	p.G4566E	43 %
P25	POT1	25913	NM_015450.2	7	124290902	Missense	c.284G>T	p.G95V	41 %
P25	HMCN1	83872	NM_031935.2	1	184275588	Missense	c.6134G>T	p.W2045L	31 %
P25	STC1	6781	NM_003155.2	8	23764905	Missense	c.346A>T	p.T116S	52 %
P25	PKP2	5318	NM_001005242.2	12	32865686	Missense	c.1884G>C	p.K628N	48 %
P25	ODZ1/ TENM1	10178	NM_014253.3	X	123342366	STOP	c.7879G>T	p.G2627*	26 %
P25	C1orf84/ SZT2	23334	NM_015284.3	1	43680313	Missense	c.7627C>T	p.P2543S	52 %
P25	VWF	7450	NM_000552.3	12	5951055	Missense	c.7519C>T	p.R2507W	42 %
P25	ZC3H18	124245	NM_144604.3	16	87205192	STOP	c.1222C>T	p.R408*	60 %
P25	CDAN1	146059	NM_138477.2	15	40809351	Missense	c.2287T>C	p.F763L	44 %
P25	C13orf26/ TEX26	122046	NM_152325.1	13	30441073	Missense	c.698A>C	p.Q233P	43 %
P25	ERC2	26059	NM_015576.1	3	56305486	Missense	c.675C>G	p.I225M	39 %
P25	CDC23	8697	NM_004661.3	5	137562054	Missense	c.890T>C	p.I297T	52 %

Table 12 A complete list of somatic nonsynonymous point mutations and one splice site mutation from 25 exomes. Var. Freq.: Variant Frequency

Case	Gene Name	Gene ID	Annotation	Chr	Position	cDNA change	Protein change
P03	NOTCH1	4851	NM_017617.3	9	138510470	c.7541_7542delCT	p.(Pro2514Argfs*4)
P04	DZIP1L	199221	NM_173543.2	3	139273257	c.1531_1533delinsCAT	p.(Lys511His)
P05	POLN	353497	NM_181808.2	4	2150892	c.1119delT	p.(Cys373Trpfs*9)
P10	RNF219	79596	NM_024546.3	13	78111006	c.503del	p.(Asn168Metfs*7)
P12	MBL2	4153	NM_000242.2	10	54200521	c.219dup	p.(Gly74Trpfs*31)
P12	MYST4/ KAT6B	23522	NM_012330.3	10	76451845	c.3231_3232insCGAG GAGGA	p.(Asp1077_Glu107 8insArgGlyGly)
P12	SPINK7	84651	NM_032566.2	5	147673148	c.87del	p.(Val30Trpfs*42)
P16	SUSD4	55061	NM_017982.3	1	221603325	c.63_65del	p.(Gln22del)
P16	MATN2	4147	NM_002380.3	8	99060403	c.1073del	p.(Thr358Serfs*5)
P21	OR4A16	81327	NM_001005274.1	11	54867315	c.70dup	p.(Thr24Asnfs*63)
P22	STAG2	10735	NM_001042749.1	X	123042954	c.2819_2828del	p.(Arg940Ilefs*6)
P22	CIB2	10518	NM_006383.3	15	76203112	c.73_75del	p.(Lys25del)
P22	DNAJC2	27000	NM_014377.1	7	102755417	c.351del	p.(Lys117Asnfs*14)

Table 13 A complete list of small somatic InDels from the 25 exomes.

Point mutations can be divided into transitions and transversions. A mutation from one purine base to the other (A->G or G->A) or one pyrimidine base to the other (C->T or T->C) is called a transition. Transversions are base substitutions between purines and pyrimidines. We calculated the frequency of transitions and transversions in our data set. Transitions from a guanine base to an adenine or from a cytosine base to a thymine base, on the minus strand, were the most frequent base substitution, accounting for about 65 % of mutations observed. Transversions, on the other hand, were less frequently detected (35 %).

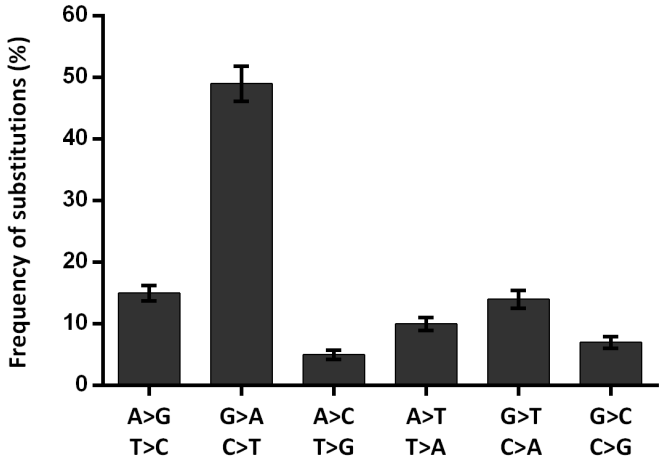


Figure 24 Frequency of substitutions detected in 25 CLL exomes. There are six different types of single base substitutions. Most frequent is the transition G>A or C>T. Error bars represent standard deviation.

The patients had about 11 (±4.3) mutations on average as detected by our exome sequencing pipeline. We only consider mutations here that lead to an amino acid change. The lowest number of mutations (n=3) was found in two patients. Both had trisomy 12. One had non-

mutated and the other mutated *IGHV* genes. The highest number of mutations with n=20 was observed in a patient who had a mutated *IGHV* status and a 13q deletion.

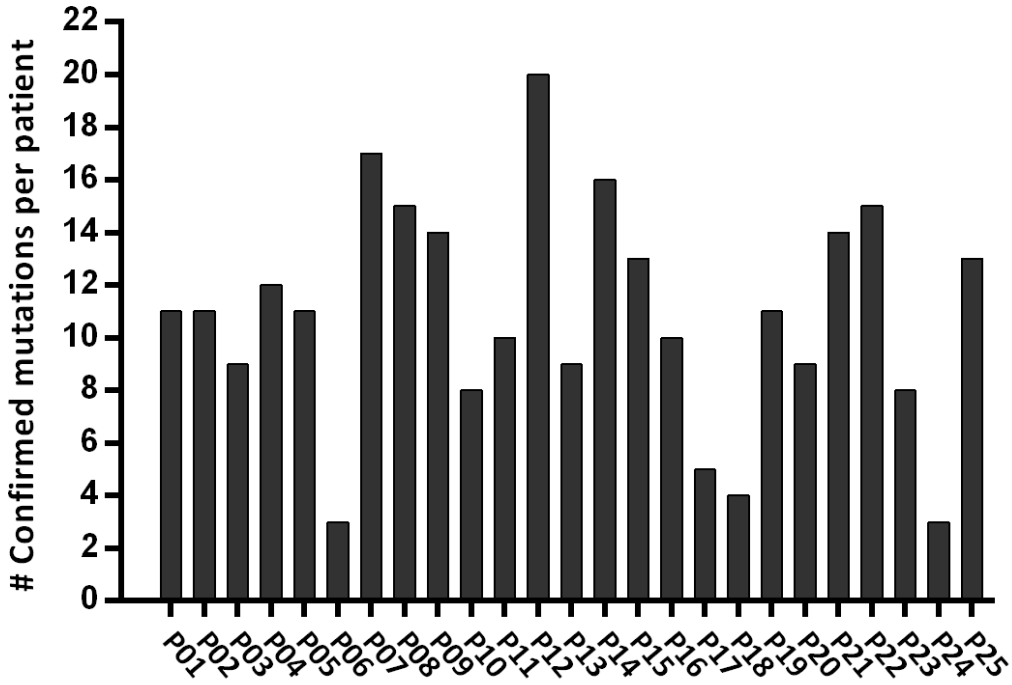


Figure 25 Total number of confirmed mutations per patient. The total number of mutations include: non-synonymous point mutations, InDels and splice site mutations

Although, we did not validate (Sanger resequence) putative mutations that would not lead to an amino acid change (synonymous mutations), we can estimate the number of true positive synonymous mutations from the confirmation rate obtained for the nonsynonymous mutations. If we use the data from the somatic workflow, we can estimate the number of true synonymous mutations using a confirmation rate of 94 % (Figure 23). Graur and Li computed the proportion of the different types of mutations from the genetic code with the assumption that all possible mutations occur with the same frequencies and that there is no codon preference (Sinauer Associates Graur and Li 2000). There are 549 possible point mutations that can occur in coding triplets (61x9). 134 or 25 % out of these 549 mutations do not lead to an amino acid change (silent or synonymous mutations). 415 (75 %) of the possible mutations lead to a change in the amino acid sequence either to a missense mutation (n=392, 71 %) or to a nonsense mutation (n=23, 4 %) (Sinauer Associates Graur and Li 2000). Based on the confirmation rate of 94 % (somatic workflow) and the fact that we detected 102 putative silent mutations in the 25 exomes (somatic workflow only) we estimate that there are about 96 true silent mutations. This is equivalent to about 27 % of all point mutations. Thus, we have approximately 73 % nonsynonymous (including 5 % nonsense mutations) and 27 % silent mutations in the coding sequences of our 25 CLL samples. This result is in good agreement with the theoretical calculations of Graur and Li and suggests that there is no obvious selection effect for nonsynonymous mutations in our 25 CLL samples.

4.1.4.1 *Mutation Frequency in CLL Samples with Mutated vs. Unmutated IGHV Genes*

Somatic hypermutation may not only affect the immunoglobulin locus but also other genes (Wang *et al.* 2004). One would therefore assume that the number of mutations in our CLL samples with mutated *IGHV* genes should be greater than in patients with an unmutated *IGHV* status.

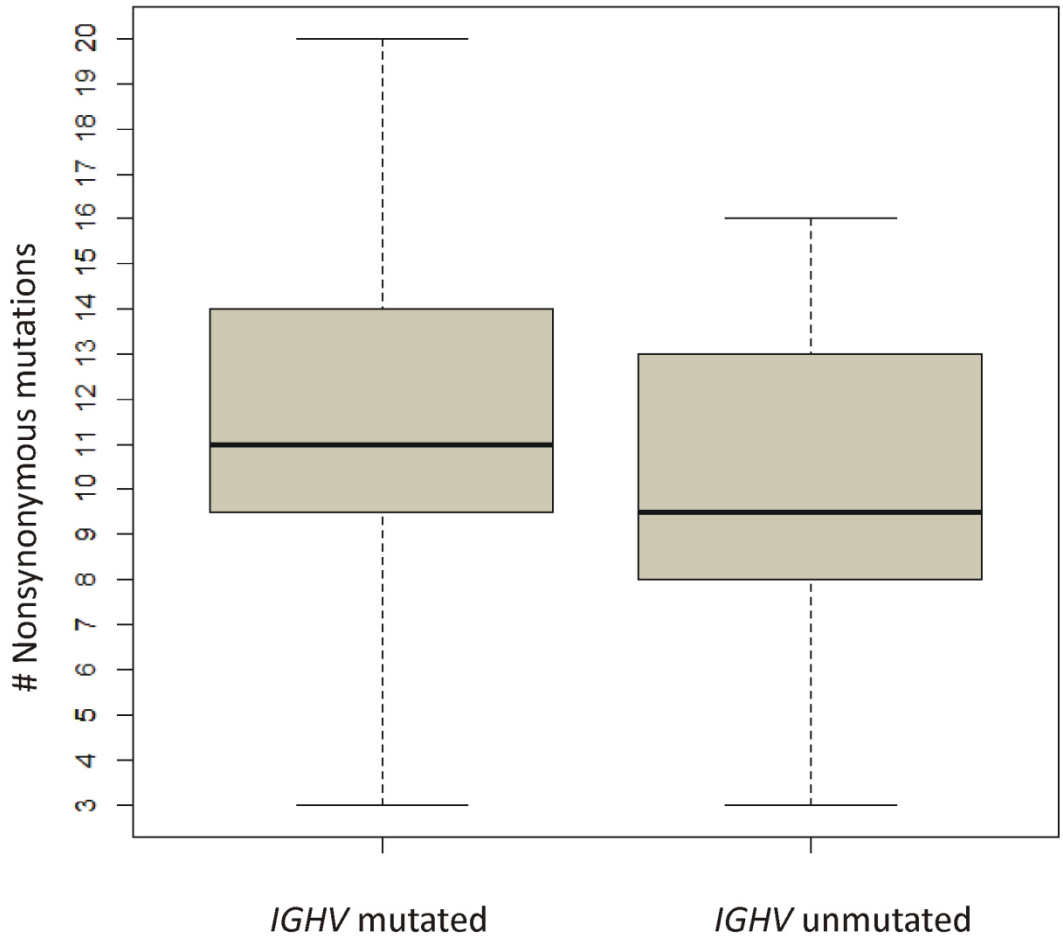


Figure 26 Box plot showing the number of nonsynonymous mutations according to *IGHV* mutational status. There was no significant difference between the two groups, Mann-Whitney U test P= 0.4864.

Although there seems to be a trend that patients with mutated *IGHV* harbor more nonsynonymous mutations (including InDels) the difference in mutation frequency did not reach statistical significance (Figure 26).

4.1.5 Discriminating Between Driver and Passenger Mutations

A major challenge of large-scale mutation screens of tumors using next-generation sequencing is to distinguish between “driver” and “passenger” mutations. Passenger mutations are somatic mutations that just happened to be in the original, tumor-initiating cell. Passenger mutations do not have any functional consequences that affect the malignant phenotype of the cancer cells. Driver mutations, on the other hand, are mutations, which alter the function of genes in such a way that the consequences of these altered gene functions lead to or enhances the malignant phenotype. It is thus, very important to determine whether a given mutation is a passenger or a driver mutations. There are basically two approaches to distinguish between driver and passenger mutations: 1) functional validation by establishing for example a cell line model system for the mutation in question or 2) determining whether a gene is recurrently mutated. Of course a functional validation, even though it is the gold standard for identifying driver mutations, is very time-consuming and not practical for large numbers of mutations. Therefore identifying recurrently mutated genes in a given cancer subtype is the method of choice; even though genes can also be recurrently mutated by chance. As a first step to identify driver mutations we looked for recurrently mutated genes in our exome dataset. To extend our search for recurrently mutated genes we also compared our datasets with results from published CLL exomes and genomes.

4.1.5.1 *Recurrently Mutated Genes*

A recurrently mutated gene is a gene found mutated in more than one patient sample.

Case	Gene	Chr	Position	cDNA	AA change	*IGHV* status	FISH status	Variant frequency
P03	*XPO1*	2	61572976	G1711A	E571K	U-CLL	normal	20 %
P04	*XPO1*	2	61572975	A1712G	E571G	U-CLL	13q	53 %
P14	*XPO1*	2	61572976	G1711A	E571K	U-CLL	13q	7 %
P15	*XPO1*	2	61572976	G1711A	E571K	U-CLL	13q	30 %
P04	*SF3B1*	2	197975079	A2098G	K700E	U-CLL	13q	36 %
P15	*SF3B1*	2	197975066	T2111A	I704E	U-CLL	13q	42 %
P20	*SF3B1*	2	197975618	C1984G	H662D	U-CLL	normal	48 %
P19	*KRAS*	12	25289551	G35A	G12D	M-CLL	normal	42 %
P22	*KRAS*	12	25289548	G38A	G13D	U-CLL	11q	52 %
P02	*CDH12*	5	22114366	T177A	N59K	M-CLL	normal	55 %
P12	*CDH12*	5	21890556	C627A	F209L	M-CLL	13q	51 %
P05	*PCLO*	7	82422083	A6122G	E2041A	M-CLL	Tris12	40 %
P15	*PCLO*	7	82419594	C8611T	P2871S	U-CLL	13q	43 %
P05	*MYH2*	17	10384070	G1047T	K349N	M-CLL	Tris12	32 %
P09	*MYH2*	17	10389453	A440G	K147R	M-CLL	13q	45 %

Table 14 Recurrently mutated genes in our 25 CLL sample cohort. MYH2: 1941 AA; PCLO: 4935 AA; CDH12: 794 AA; XPO1: 1071 AA; SF1B3: 144 AA; KRAS: 188 AA. AA: Amino Acid, M-CLL: mutated CLL, U-CLL: unmutated CLL

The larger the coding region of a gene, the higher the likelihood it is recurrently mutated by chance. The most frequently mutated gene in our patient cohort was *XPO1*. *XPO1* (exportin 1), also known as *CRM1*, encodes a 1071 amino acid long protein that mediates nuclear export signal (NES)-dependent protein export of proteins from the nucleus into the cytoplasm. *XPO1* was found mutated in 16 % (4/25) of our samples. One of these mutations was not initially detected by our exome sequencing analysis pipeline but only when we performed a larger Sanger sequencing screen of our cohort (4.2). The mutation evaded detection in the exome analysis pipeline due to a low variant frequency of just 7 % in the CLL sample. In Sanger sequencing a very small clone was detected that could retrospectively also be found in the next-generation sequencing data.

As an example, the XPO1 mutation of patient P04 is shown in Figure 27. The grey bars represent reads aligned to the reference sequence. In the upper panel, representing the CLL sample, the variant at position 61,572,976 on chromosome 2 causing the amino acid change p.E571G is shown as blue Cs (Figure 27 A). Amino acid E571 is a mutational hotspot since it was mutated in all 4 of our samples. In the remission sample, in the lower panel, no variant was detected. Variant frequencies and total read counts are displayed in the small boxes. In addition, the number of reads derived from the plus and minus strand is indicated by a plus and minus sign, respectively. To confirm the result of the exome sequencing pipeline, the region was analyzed by Sanger sequencing (Figure 27 B). The upper chromatogram is derived from the CLL with the *XPO1* mutation and the lower one is derived from the remission sample, which is wild type.

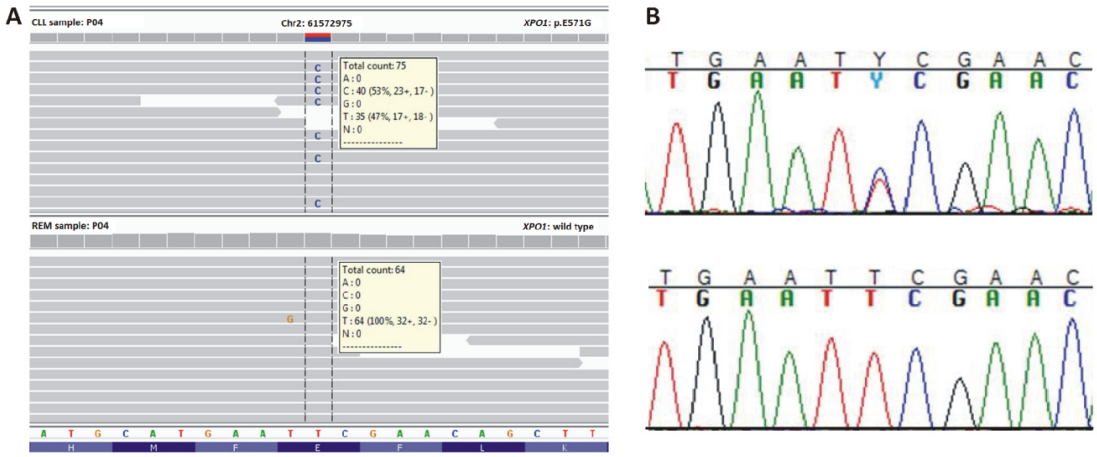


Figure 27 *XPO1* mutation in patient P04. A) Exome data shown as an IGV screenshot. The upper panel represents the CLL sample with the heterozygous mutation. No variant was detected in the remission sample in the lower panel. The variant counts are displayed in small boxes. B) Sanger sequencing of the confirmed mutation in the CLL sample (upper panel). In the remission sample (lower panel) the hotspot was wild type.

Another gene, which was found to be frequently mutated in our CLL samples as well as in other hematological disorders (Papaemmanuil *et al.* 2011), is *SF3B1,* a gene which encodes the subunit 1 of the splicing factor 3b protein complex. With a frequency of 12 %, *SF3B1* is the second most frequently mutated gene in our 25 CLL samples. In contrast to *XPO1,* the mutations in *SF3B1* are not limited to a hotspot. Mutations in the *CDH12* gene were also found

with a frequency of 8 %. This gene belongs to the cadherin superfamily and is involved in cell-cell adhesion. We also detected mutations in the KRAS hotspot at amino acid 12 and 13 in 2/25 cases (8 %). This protein is a member of the small GTPase superfamily. PCLO plays a role in synaptic vesicle trafficking. It was found mutated in 2/25 (8 %) of our CLL patients. We also found *MYH2* mutations in 2/25 cases (8 %), this gene encodes the myosin 2 motor protein. It plays an important role in the contraction of skeletal muscles.

4.1.5.2 *Comparison of Mutations found in the Study Cohort with Published CLL Mutation Data*

To identify additional recurrently mutated genes, we compared the mutations we found in the exomes of our 25 CLL samples with published CLL whole exome and whole genome sequencing datasets. A total of 2399 genes with non-silent mutations affecting their protein coding regions have been described in three studies examining 200 CLL patient samples in total. The patients in these studies were predominantly analyzed at the time before their first treatment (Puente *et al.* 2011; Wang *et al.* 2011; Quesada *et al.* 2012). We compared the non-silent mutations in coding regions of the published datasets with the non-silent coding region mutations of our own dataset (n=270 mutations) which were found in 259 different genes.

	Wang	**Quesada**	**Puente**	**Wang Quesada Puente**	**REACH**
Total number of non-silent coding mutations	1730	1165	45	2935	270
Total number of mutated genes	1521	1027	44	2399	259
Recurrently mutated genes: mutated in 2 or more sample	139	96	0	194 (183)	6
Genes which have two mutations in the same patient sample	7	8	1	-	2

Table 15 Summary of non-silent coding region mutations found in our study cohort and published CLL datasets. A recurrently mutated gene is a gene that was found mutated in more than one sample. A gene that is found mutated more than once but in the same sample is not considered as a recurrently mutated gene (see last row of the table). All three public data sets share 183 genes, 194 were recurrently mutated when mutated genes are added that were recurrently mutated in only one of the public datasets.

Wang and colleagues analyzed n=91 patient samples with whole exome and whole genome sequencing and detected 1730 non-silent coding region mutations in 1521 genes, 139 genes were found recurrently mutated. Quesada and colleagues analyzed 105 patient and found 1165 non-silent coding region mutations in 1027 different genes. Of these genes, 96 were recurrently mutated. By whole genome sequencing Puente *et al* detected 45 non-silent coding region mutations in 44 genes in four patient samples. None of them were recurrently mutated, the only mutation that occurred twice within the same gene was found in the same sample. When we consider all three studies, 194 of the 2399 genes were found mutated in more than

one CLL sample. All three studies together have 183 mutated genes that were found in more than one study. Ten genes were found mutated in all three studies (Figure 28 A).

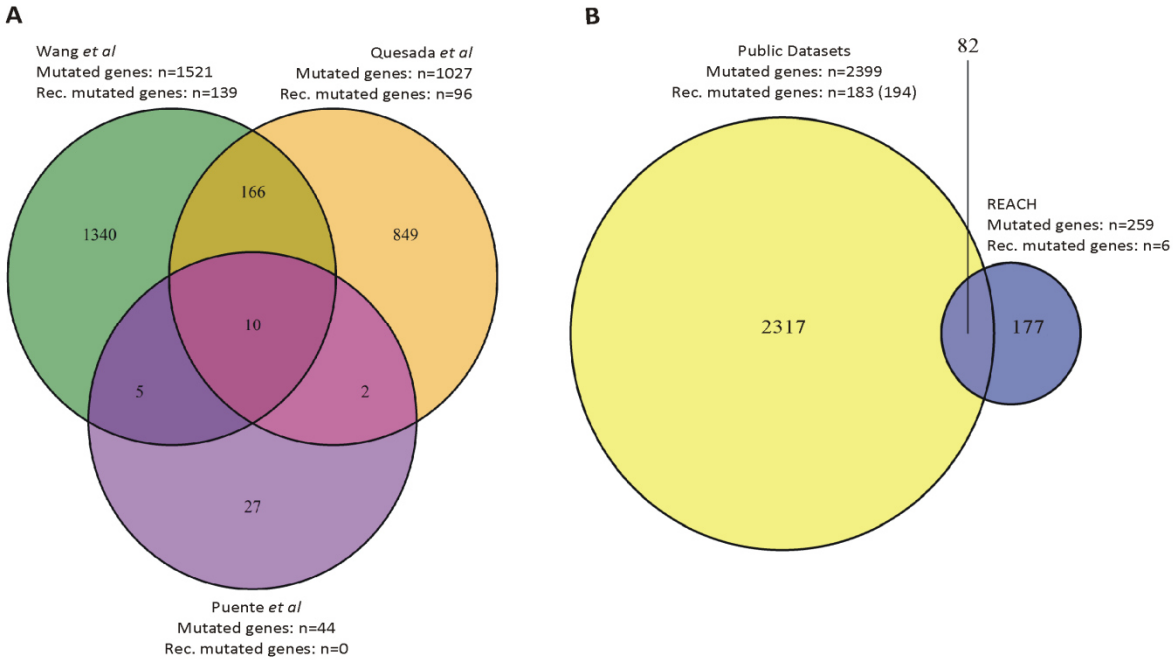


Figure 28 Mutated genes in CLL: Comparison of publicly available data sets with the data from our study. A) We selected three large CLL exome and genome sequencing studies: Wang *et al* (green), Quesada *et al* (orange) and Puente *et al* (purple). In these three studies n=91, n=105 and n=4 patients were examined, identifying n=1521, n=1027, and n=44 different genes with non-silent mutations in their coding regions, respectively. The Venn diagram shows the number of genes that were found to be mutated and the number of identical genes found in the three studies. B) When we compared our dataset of n=259 mutated genes (blue circle) with all mutated genes from the three studies (n=2399) there were 82 genes in common.

When we compared the 2399 genes found mutated in these three studies with our set of 259 mutated genes we found an overlap of 82 genes. Thus, our study (REACH) reports 177 different genes that have been found mutated in CLL for the first time. Of the 82 genes found mutated both in the three studies and our study, 47 can only be identified as recurrently mutated when our dataset is taken into consideration. Thus, our study increases the number of recurrently mutated genes in CLL from 194 to 241. The 47 genes that our efforts add to the list of recurrently mutated genes and their full name are listed below.

Genes only identified as recurrently mutated when our dataset is taken into consideration.

Gene ID	Official Symbol	Official Full Name (HGNC)
338	APOB	apolipoprotein B
1010	CDH12	cadherin 12, type 2 (N-cadherin 2)
1109	AKR1C4	aldo-keto reductase family 1, member C4
1290	COL5A2	collagen, type V, alpha 2
1301	COL11A1	collagen, type XI, alpha 1
1387	CREBBP	CREB binding protein
1734	DIO2	deiodinase, iodothyronine, type II
1767	DNAH5	dynein, axonemal, heavy chain 5
1793	DOCK1	dedicator of cytokinesis 1
1832	DSP	desmoplakin
2317	FLNB	filamin B, beta
2318	FLNC	filamin C, gamma
3762	KCNJ5	potassium inwardly-rectifying channel, subfamily J, member 5
3845	KRAS	Kirsten rat sarcoma viral oncogene homolog
4620	MYH2	myosin, heavy chain 2, skeletal muscle, adult
4893	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog
5138	PDE2A	Phosphodiesterase 2A cGMP-stimulated
6336	SCN10A	sodium channel, voltage-gated, type X, alpha subunit
6903	TBCC	tubulin folding cofactor C
7369	UMOD	Uromodulin
8549	LGR5	leucine-rich repeat containing G protein-coupled receptor 5
10076	PTPRU	protein tyrosine phosphatase, receptor type, U
10922	FASTK	Fas-activated serine/threonine kinase
11093	ADAMTS13	ADAM metallopeptidase with thrombospondin type 1 motif, 13
11122	PTPRT	protein tyrosine phosphatase, receptor type, T
11177	BAZ1A	bromodomain adjacent to zinc finger domain, 1A
23036	ZNF292	zinc finger protein 292
23334	SZT2	seizure threshold 2 homolog (mouse)
23517	SKIV2L2	superkiller viralicidic activity 2 –like (*S. cerevisiae*)
26059	ERC2	ELKS/RAB6-interacting/CAST family member 2
26999	CYFIP2	Cytoplasmic FMR1 interacting protein 2
64759	TNS3	tensin 3
80070	ADAMTS20	ADAM metallopeptidase with thrombospondin type 1 motif, 20
83873	GPR61	G protein-coupled receptor 61
89857	KLHL6	kelch-like family member 6
91156	IGFN1	immunoglobulin-like and fibronectin type III domain containing 1
114800	CCDC85A	coiled-coil domain containing 85A
128553	TSHZ2	teashirt zinc finger homeobox 2
131890	GRK7	G protein-coupled receptor kinase 7
139135	PASD1	PAS domain containing 1
165186	FAM179A	family with sequence similarity 179, member A
199221	DZIP1L	DAZ interacting zinc finger protein 1-like

Genes only identified as recurrently mutated when our dataset is taken into consideration.		
Gene ID	**Official Symbol**	**Official Full Name (HGNC)**
282775	OR5J2	olfactory receptor, family 5, subfamily J, member 2
283450	HECTD4	HECT domain containing E3 ubiquitin protein ligase 4
339976	TRIML1	tripartite motif family-like 1
340533	KIAA2022	KIAA2022
347273	MURC	muscle-related coiled-coil protein

Table 16 List of 47 recurrently mutated genes. With these genes our study increases the number of recurrently mutated genes in CLL from 194 to 241. HGNC: HUGO Gene Nomenclature Committee

From the 82 genes that overlap between the public datasets and our dataset, 47 could be newly added to the recurrently mutated genes through the analysis of our CLL cohort (Table 16). The remaining 35 genes were already known to be recurrently mutated from the three public datasets. Together with our dataset these genes are mutated in at least 3 out of 225 patients (25 patients from our dataset and 200 patients from the three public datasets).

Genes found mutated in our dataset and also known to be recurrently mutated in the public datasets				
Gene ID	**Official Symbol**	**Official full name (HGNC)**	**Cases/225**	**Frequency**
23451	SF3B1	splicing factor 3b, subunit 1, 155kDa	27/225	12 %
7157	TP53	tumor protein p53	14/225	6 %
4851	NOTCH1	notch 1	11/225	5 %
472	ATM	ataxia telangiectasia mutated	10/225	4 %
27445	PCLO	piccolo presynaptic cytomatrix protein	9/225	4 %
7514	XPO1	exportin 1 (CRM1 homolog, yeast)	8/225	4 %
667	DST	Dystonin	7/225	3 %
64478	CSMD1	CUB and Sushi multiple domains 1	7/225	3 %
1106	CHD2	chromodomain helicase DNA binding protein 2	6/225	3 %
9968	MED12	mediator complex subunit 12	6/225	3 %
25913	POT1	protection of telomeres 1	6/225	3 %
53353	LRP1B	low density lipoprotein receptor-related protein 1B	6/225	3 %
79633	FAT4	FAT atypical cadherin 4	5/225	2 %
1654	DDX3X	DEAD (Asp-Glu-Ala-Asp) box helicase 3, X-linked	5/225	2 %
1959	EGR2	early growth response 2	5/225	2 %
7273	TTN	Titin	5/225	2 %
120114	FAT3	FAT atypical cadherin 3	5/225	2 %
6091	ROBO1	roundabout, axon guidance receptor, homolog 1 (Drosophila)	4/225	2 %
8880	FUBP1	far upstream element (FUSE) binding protein 1	4/225	2 %
9152	SLC6A5	solute carrier family 6	4/225	2 %

Genes found mutated in our dataset and also known to be recurrently mutated in the public datasets				
Gene ID	**Official Symbol**	**Official full name (HGNC)**	**Cases/225**	**Frequency**
		(neurotransmitter transporter), member 5		
9353	SLIT2	slit homolog 2 (Drosophila)	4/225	2 %
9699	RIMS2	regulating synaptic membrane exocytosis 2	4/225	2 %
9732	DOCK4	dedicator of cytokinesis 4	4/225	2 %
1285	COL4A3	collagen, type IV, alpha 3	3/225	1 %
5137	PDE1C	phosphodiesterase 1C, calmodulin-dependent 70kDa	3/225	1 %
7450	VWF	von Willebrand factor	3/225	1 %
10082	GPC6	glypican 6	3/225	1 %
10178	TENM1/ ODZ1	teneurin transmembrane protein 1	3/225	1 %
23347	SMCHD1/ AP001011 .3	structural maintenance of chromosomes flexible hinge domain containing 1	3/225	1 %
55061	SUSD4	sushi domain containing 4	3/225	1 %
83872	HMCN1	hemicentin 1	3/225	1 %
84033	OBSCN	obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	3/225	1 %
124245	ZC3H18	Zink finger CCH-type containing 18	3/225	1 %
285175	UNC80/ C2orf21	Unc-80 homolog (*C.elegans*)	3/225	1 %
440279	UNC13C	Unc-13 homolog c (*C.elegans*)	3/225	1 %

Table 17 List of 35 recurrent genes. Genes that were found mutated in our gene set but were already known to be recurrently mutated within the three public datasets. HGNC: HUGO Gene Nomenclature Committee

4.1.5.3 *Coverage of Frequently Mutated Genes in the CLL Exome Study Cohort*

SF3B1, ATM, MYD88, TP53 and *NOTCH1* are among the most frequently mutated genes in CLL. However, we did not find a mutation in *MYD88*. To judge how well these genes were enriched by Agilent's SureSelect 50 Mb capture kit we calculated the coverage per nucleotide of these five genes. We also calculated the per nucleotide coverage of *XPO1*, *KRAS*, *PCLO*, *CDH12* and *MYH2* which were also found recurrently mutated in our 25 CLL samples (Table 14). *CDH12* is the only gene that had not been described as recurrently mutated in the three published datasets.

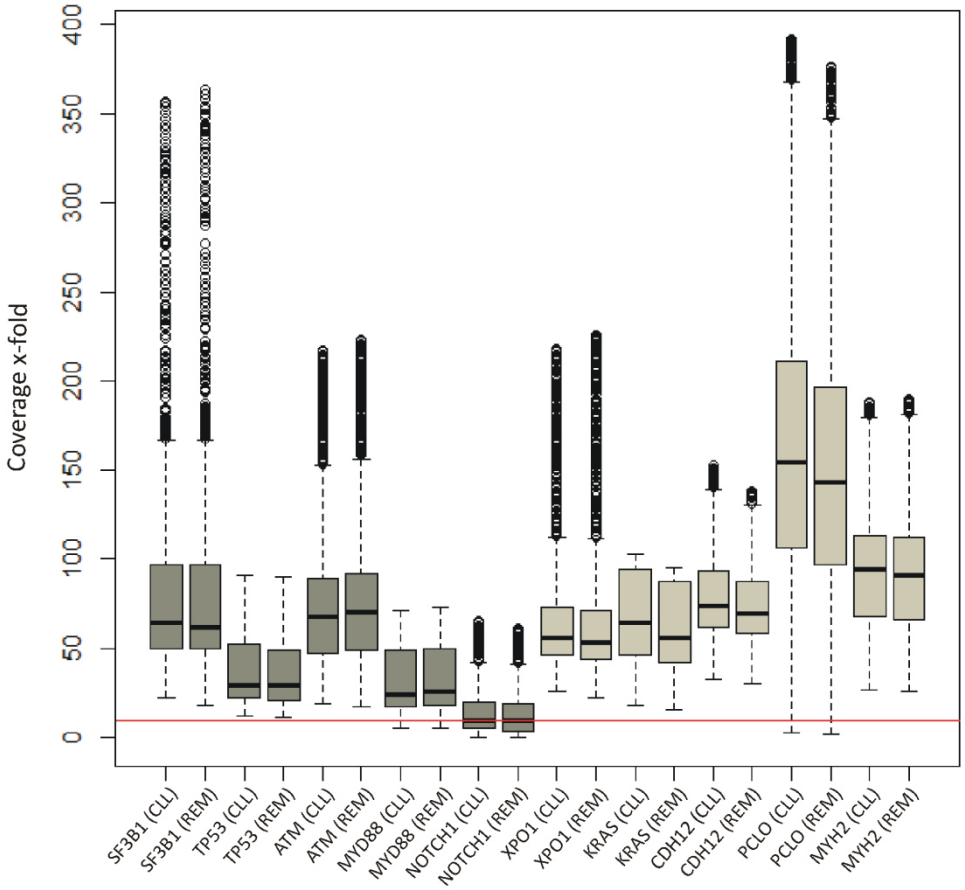


Figure 29 Average coverage per nucleotide position of frequently mutated target genes in our CLL cohort. *SF3B1*, *ATM*, *MYD88*, *TP53* and *NOTCH1* are frequently mutated genes in CLL. *XPO1*, *KRAS*, *CDH12*, *PCLO* and *MYH2* were found to be recurrently mutated in our study. CLL: denotes the reads in the leukemia samples; REM: denotes the reads in the remission samples

Except for *NOTCH1*, all nucleotides in these 10 genes have a minimum coverage of ten reads, which is sufficient for SNV calling. Approximately half of all positions of *NOTCH1* did not reach the minimum requirement of 10 reads in the tumor or remission samples. With a CG content of >60 %, the *NOTCH1* cDNA has the highest GC content of all these genes. It should be noted that in some samples the mutations in more than one of these genes occur together. Patient P15 and P04 harbor both a *XPO1* and a *SF3B1* mutation and P03 has a *XPO1* and a *NOTCH1* mutation.

4.1.5.4 *Gene Category Analysis of Mutated Genes*

We identified mutations in 259 genes in our 25 CLL samples. As explained earlier, it is very difficult to distinguish driver from passenger mutations. There were 82 genes in our 25 samples that were recurrently mutated if we take into consideration the whole exome and whole genome sequencing data from a total of 225 samples which include 200 CLL samples described in the literature. These 82 genes are quite likely to harbor driver mutations. However, there are 177 genes in our 25 CLL samples, which were only found mutated once. Not all of these genes would harbor driver mutations but some of them might. We would assume that driver mutations preferentially affect genes involved in certain cellular pathways. To estimate whether there was a preference for mutated genes to belong to certain groups we compared our mutated gene list to the Molecular Signature Database, a collection of annotated gene sets (MSigDB v.3.1) (Subramanian, *et al*. PNAS 2005). User-defined lists can be compared with the database and overlaps can be computed (Liberzon *et al.* 2011). In this way, it can be possible to find, that a large gene set contains mainly genes that are involved in a limited number of pathways.

Table 18 REACH dataset clustered by gene families as analyzed by the MSigDB. Using the online tool at http://www.broadinstitute.org/gsea/msigdb/annotate.jsp we categorized the 259 mutated genes from our 25 CLL samples. All genes of the MSigDB are categorized into eight different gene categories, which are relevant in tumorigenesis. We loaded the list of 259 genes that we found mutated (coding mutations) in our own dataset into the search mask of MSigDB v.3.1 in order to compare the REACH dataset with the MsigDB v.3.1 database and categorize the genes by the given gene families. The output of this comparison is a table containing all genes of our own dataset that overlap with the MsigDB genes and could be grouped into one or more of the eight gene categories. A gene can also belong to more than one gene category at the same time. The eight different gene categories are listed within the grey boxes of the first row and also the first column of the table. Within the green boxes we find all genes belonging to one gene category. If a gene is also listed in one or more of the white boxes above, it belongs to more than one category. The grey boxes that define the position of the white boxes, tell us which category the gene belongs to. From this table one can see that the gene *TP53* for example is a member of the tumor suppressor but it is also a member of the transcription factor gene category. The *CREBBP* is a transcription factor, translocated cancer gene and oncogene at the same time.

	Cytokines and Growth Factors	Transcription Factors	Homeodomain Proteins	Cell Differentiation Markers	Protein Kinases	Translocated Cancer Genes	Oncogenes	Tumor Suppressors
Tumor Suppressors	-	*TP53*	-	-	*ATM*	-	-	*ATM, MSH6, SETD2, TNFAIP3, TP53*
Oncogenes	-	*CREBBP*	-	-	-	*CREBBP, NOTCH1*	*CREBBP, KRAS, NOTCH1, NRAS*	
Translocated Cancer Genes	-	*CREBBP*	-	-	-	*CREBBP, NOTCH1*		
Protein Kinases	-	*BRDT*	-	-	*AMHR2, ATM, BMP2K, BRDT, FASTK, GRK7, NRBP1,OBSCN, PRKCA, RPS6KA3, TESK2, TTN*			
Cell Differentiation Markers	-	-	-	*ADAM17, CSF2RB, CXCR4, EMR2, IL3RA, KLRC2*				
Homeodomain Proteins	-	*TSHZ2*	*TSHZ2*					
Transcription Factors	-	*ASH1L, BAZ1A, BRDT, CHD2, CREBBP, EGR2, FUBP1, KAT5, MED12, OTUD7B, PACS2,TP53, TSHZ2, ZFP161*						
Cytokines and Growth Factors	*FGF14, IL28A, RNASE2, SLIT2, STC1, VEGFC*							

Patient	Gene Name	Accession Number	Transcript Change	Protein Change
P1	CHD2	NM_001271.3	c.3455G>A	p.(Arg1152Gln)
P1	IL28A	NM_172138.1	c.280C>T	p.(Arg94Cys)
P2	CREBBP	NM_004380.2	c.46A>T	p.(Lys16*)
P3	RPS6KA3	NM_004586.2	c.506A>G	p.(Asp169Gly)
P3	NOTCH1	NM_017617.3	c.7541_7542delCT	p.(Pro2514Argfs*4)
P3	MED12	NM_005120.2	c.131G>A	p.(Gly44Asp)
P4	EMR2	NM_013447.3	c.283G>A	p.(Val95Met)
P4	CSF2RB	NM_000395.2	c.2425C>G	p.(Gln809Glu)
P5	KAT5	NM_006388.3	c.1018G>A	p.(Val340Ile)
P5	BRDT	NM_207189.2	c.1934G>A	p.(Ser645Asn)
P6	GRK7	NM_139209.2	c.941A>G	p.(Tyr314Cys)
P7	BAZ1A	NM_013448.2	c.2995A>G	p.(Lys999Glu)
P7	OTUD7B	NM_020205.2	c.2051G>T	p.(Arg684Met)
P7	BMP2K	NM_198892.1	c.1617G>T	p.(Gln539His)
P8	TTN	NM_133378.4	c.58999G>A	p.(Val19667Ile)
P8	TNFAIP3	NM_006290.3	c.90T>A	p.(Phe30Leu)
P9	ASH1L	NM_018489.2	c.4868G>A	p.(Gly1623Asp)
P9	AMHR2	NM_020547.2	c.137G>T	p.(Gly46Val)
P10	PRKCA	NM_002737.2	c.1439C>T	p.(Ala480Val)
P10	KLRC2	NM_002260.3	c.358G>A	p.(Glu120Lys)
P11	ZFP161	NM_003409.4	c.314T>A	p.(Val105Asp)
P11	NRBP1	NM_013392.2	c.169G>T	p.(Glu57*)
P11	ATM	NM_000051.3	c.9023G>A	p.(Arg3008His)
P11	EGR2	NM_000399.3	c.1066G>A	p.(Glu356Lys)
P12	RNASE2	NM_002934.2	c.334C>T	p.(Leu112Phe)
P12	ADAM17	NM_003183.4	c.1481T>G	p.(Met494Arg)
P12	PACS2	NM_015197.3	c.448G>A	p.(Gly150Ser)
P14	FUBP1	NM_003902.3	c.41C>G	p.(Ser14*)
P15	SLIT2	NM_004787.1	c.199A>T	p.(Ile67Phe)
P16	OBSCN	NM_001271223.2	c.18892G>T	p.(Ala6298Ser)
P19	FGF14	NM_175929.2	c.634C>T	p.(Arg212*)
P19	IL3RA	NM_002183.3	c.565G>A	p.(Ala189Thr)
P19	KRAS	NM_004985.3	c.35G>A	p.(Gly12Asp)

Patient	Gene Name	Accession Number	Transcript Change	Protein Change
P21	VEGFC	NM_005429.2	c.1030A>C	p.(Lys344Gln)
P21	TESK2	NM_007170.2	c.49G>T	p.(Glu17*)
P21	TP53	NM_000546.5	c.578A>G	p.(His193Arg)
P21	FASTK	NM_033015.3	c.386C>T	p.(Thr129Met)
P21	CXCR4	NM_003467.2	c.1013C>G	p.(Ser338*)
P22	TSHZ2	NM_001193421.1	c.655G>A	p.(Ala219Thr)
P22	SETD2	NM_014159.6	c.5020G>T	p.(Glu1674*)
P23	NRAS	NM_002524.4	c.38G>T	p.(Gly13Val)
P24	MSH6	NM_000179.2	c.3604A>T	p.(Met1202Leu)
P25	STC1	NM_003155.2	c.346A>T	p.(Thr116Ser)

Table 19 MSigDB Output. The 259 mutated genes were grouped into the following categories: tumor suppressors, oncogenes, translocated cancer genes, protein kinases, cell differentiation markers, homeodomain proteins, transcritption factors and cytokines/growth factors. 21 patient samples of our dataset had at least one mutation in a gene that could be grouped to one of these gene categories which are relevant for cancer development.

We attempted to assign the 259 mutated genes (coding mutations) from our 25 CLL samples to gene categories that are known to play a critical role in cancer using MSigDB. The genes were grouped into eight categories: tumor suppressors, oncogenes, translocated cancer genes, protein kinases, cell differentiation markers, homeodomain proteins, transcription factors and cytokines/growth factors (Table 18). 43 of the 259 mutated genes belong to one of these categories. Interestingly, 21 of the 25 patients (84 %) had at least one mutation in one of these 43 genes (Table 19

4.2 *XPO1* Mutation Screening of 445 CLL Samples

We paid special attention to Exportin 1 (*XPO1*), which was the most frequently mutated gene in our patient cohort. Using exome sequencing, we initially detected three samples with mutations affecting amino acid E571 of *XPO1* in the 25 CLL samples of this study. All three *XPO1* mutated samples had a non-mutated *IGHV* status, two had a 13q deletion and one had no aberration in our FISH screen. We then screened 445 CLL samples from the REACH cohort for XPO1 mutations using Sanger sequencing of almost the entire exon 16 (amino acids 529 to 575). The 25 patient samples analyzed by whole exome sequencing were included in these 445 samples.

XPO1 Mutation	any	p.E571K	p.E571G	p.E571V	p.E571Q	p.E565I
Cases N=445	41	33	4	2	1	1
Frequency	9.2 %	7.4 %				

Table 20 *XPO1* mutations in n=445 REACH patient samples. Percentage not calculated if n<5. Note that all except one mutation affect codon 571.

The analysis of these patients detected missense mutations in residue 571 in 40 (9 %) of the 445 cases. One patient had a mutation in codon 565. Thus, the overall frequency of *XPO1* mutations is 9.2 % (41/445) in our CLL patients, who had all experienced a relapse. Among the mutated samples one was from the 25 samples of the exome sequencing group. This *XPO1* mutation, which was found in P14, was initially not detected by our exome sequencing and analysis pipeline as the number of mutated reads was only 7 %, which is below the detection threshold set at 20 %.

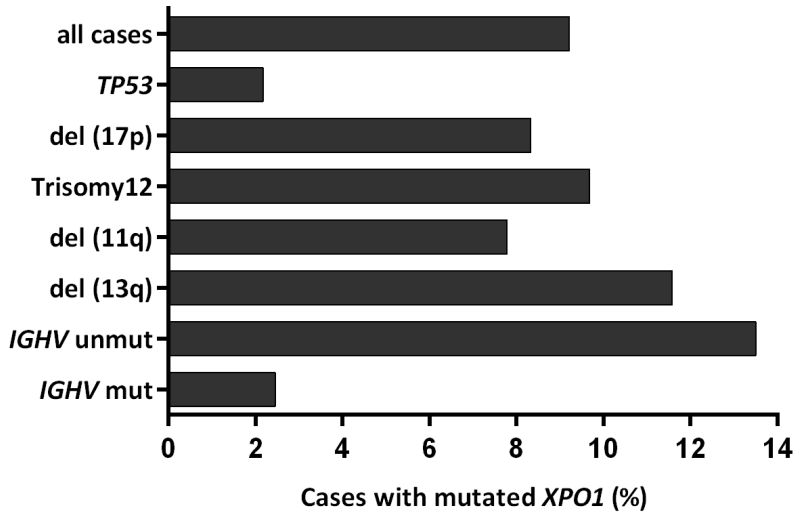


Figure 30 Frequency of *XPO1* mutations. Frequency of *XPO1* mutations in CLL subgroups defined by FISH, *IGHV* or *TP53* mutational status. Please note that these groups are not mutually exclusive. Thus the percentages add up to more than more than 9.2 %.

XPO1 mutations were slightly more frequent in patients with an unmutated *IGHV* status (13.5 %) and in patients with a 13q deletion (11.6 %).

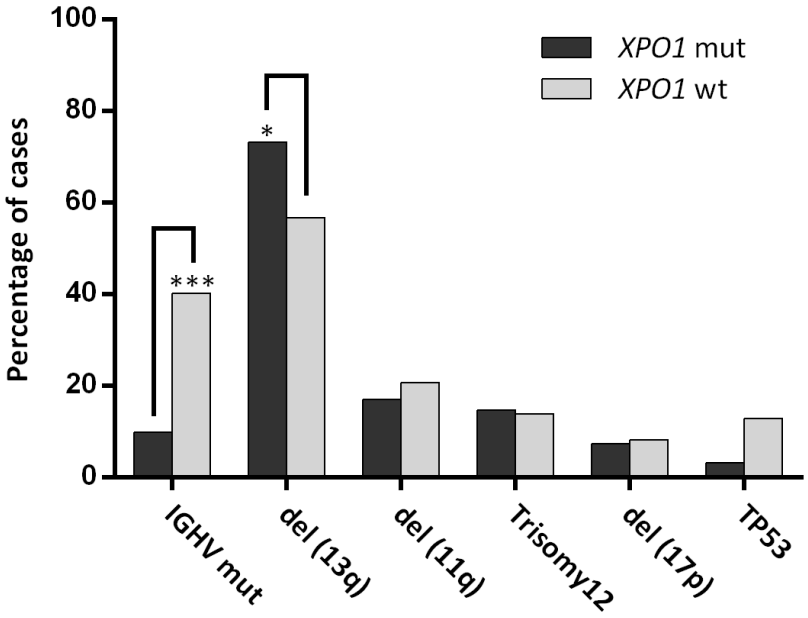


Figure 31 Distribution of *XPO1* mutations. Distribution of *IGHV* mutational status, chromosomal aberrations detected by FISH and *TP53* mutational status in previously treated CLL samples with wild type or mutant *XPO1* (Fisher's exact test: ***, P <0.0001; *, P <0.05).

In the group of 41 patients with an *XPO1* mutation, 9.8 % had a mutated *IGHV* status, 73.2 % a 13q deletion, 17.1 % an 11q deletion, 14.6 % a Trisomy 12, 7.3 % a 17p deletion and 3.1 % a mutated *TP53* gene (one patient).

In the group of patients with wild type *XPO1* (for the exon 16 region) 40.2 % had a mutated *IGHV* status, 56.7 % a 13q deletion, 20.6 % an 11q deletion, 13.9 % a Trisomy 12, 8.2 % a 17p deletion and 12.9 % a mutated *TP53* gene. Please note that these patient groups are not mutually exclusive, so the percentages add up to more than 100 %. Thus patients with an *XPO1* mutation were more likely to have a 13q deletion and less likely to have mutations of the *IGHV* genes than patients with a wild type *XPO1* gene.

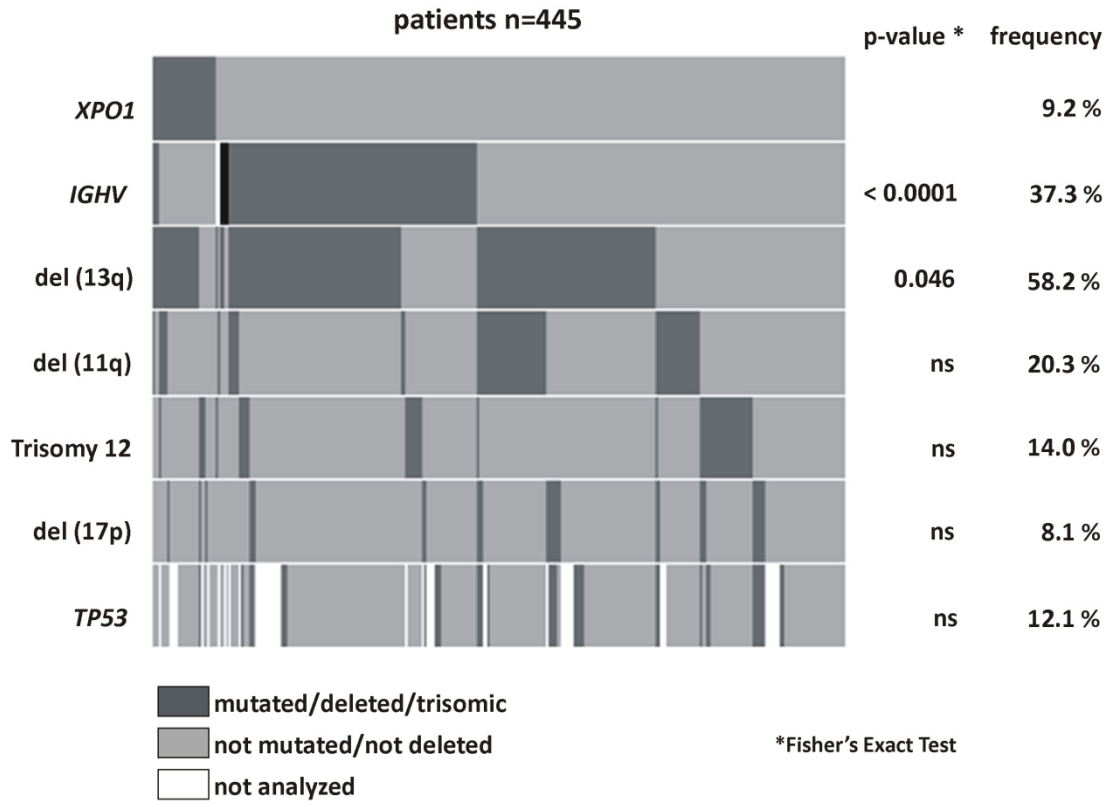


Figure 32 Heat map of the 445 CLL patients analyzed for *XPO1* mutations and 6 other genetic alterations. The p-values were calculated for the correlation of the *XPO1* mutational status with the *IGHV* mutation status, del(13q), del(11q), trisomy 12, del(17p), and *TP53* mutation. For patient samples colored in black the *IGHV* status could not be clearly determined. ns: not significant

The heat map in Figure 32 shows that mutations in *XPO1* are significantly associated with a wild type *IGHV* status and the presence of 13q deletions. The association with an unmutated *IGHV* status suggests that the *XPO1* mutations are acquired during early B-Cell development. Overall survival (OS) and progression free survival (PFS) of patients treated with FC or FC-R was kindly analyzed by F. Hoffmann-La Roche. There was no difference in OS or PFS between patients with wild type versus mutated *XPO1* (data for OS, shown in discussion). Therefore, *XPO1* is neither a prognostic nor a predictive marker for patients with previously treated CLL treated with FC or FC-R.

5 Discussion

5.1 Next Generation Sequencing and its Applications

Next generation sequencing has tremendously enhanced our ability to analyze cancer genomes and will help to improve cancer diagnostics and treatment decisions. Classical cytogenetics, a commonly used method in leukemia diagnostics, can detect chromosomal aberrations only down to about 10 Mb. FISH is more sensitive and aberrations of about 50 kb can be detected, but the targeted regions must be known and selected in advance. Array-based comparative genomic hybridization (Array-CGH) techniques allow the detection of deletions and duplications across the entire genome at a higher resolution than FISH. However, balanced translocations cannot be detected using Array-CGH technology. In contrast, NGS technology allows the analysis of a whole genome at single base pair resolution. Roche's 454 pyrosequencing technology, ABI's SOLID ligation based sequencing and Illumina's sequencing by synthesis are leading technologies in the field of next generations sequencing. These technologies enable us to study point mutations, insertions, deletions, copy number alterations or translocations in tumor genomes on a genome wide scale. WGS is the most comprehensive strategy to detect variants in tumor genomes but it is very expensive, and variants in non-coding regions, which constitute 98.5 % of the genome, are very difficult to interpret. Several whole cancer genomes have been reported so far. The first cancer genome published in 2008 was the genome of an AML patient (Ley *et al.* 2008). However, WES, which focuses on the 1.5 % of the genome coding for proteins, became more common within the last years as it is cheaper and faster than WGS. Using WES a much higher coverage of coding regions can be obtained with a considerable reduction of raw sequence production.

Initially, the targeted sequencing of coding region of a genome was only possible by transcriptome analysis of cDNA derived from mRNA. Despite the tissue specific expression patterns of genes, mutation detection in many genes is possible using transcriptome sequencing (Greif *et al.* 2011). In contrast to WES or WGS, transcriptome sequencing provides additional information on gene expression levels, alternative transcripts and fusion transcripts. However, we were unable to obtain satisfactory results using whole transcriptome sequencing for CLL samples. We found that the RNA quality was inadequate resulting in poor transcriptome coverage and a high number of duplicated reads (data not shown). This problem might be due to the fragility of CLL cells (Macdonald *et al.* 2003) which might be the cause of poor mRNA quality.

5.1.1 Whole Exome Sequencing of 25 Previously Treated CLL Patients

In 2010, Agilent Technologies introduced the first commercial whole exome capture kit, which is based on an in-solution-hybrid-selection technology (Gnirke *et al.* 2009; Bainbridge *et al.* 2010). DNA fragments containing exonic sequences are captured with biotinylated RNA baits.

We used the Agilent Human All Exon 50Mb platform to analyze a total of 50 exomes from 25 CLL patient samples. For each patient, the exome from a blood sample at diagnosis (the CLL sample) and the exome from a blood sample at remission was captured and sequenced. As exome sequencing is based on genomic DNA, each gene is evenly represented. However, a bias can be introduced due to different capturing efficiencies across the exons. The patients we analyzed were treated on a clinical study designed and sponsored by F.Hoffmann-La Roche. In this trial, previously treated patients with CD20-positive B-cell CLL who had relapsed were treated with fludarabine, a purine analog, and cyclophosphamide, an alkylating agent, alone (FC arm) or together with rituximab (R-FC arm) to evaluate the efficacy and safety of the monoclonal antibody treatment in this setting (for further details see Roche protocol BO17072). Relevant prognostic markers in CLL, like the deletion of certain genomic regions and, the mutational status of the *IGHV* region, were evaluated in the laboratory for leukemia diagnostics at the University of Munich, Campus Grosshadern. We selected 25 of 552 patient samples at the time of relapse before second-line treatment and the corresponding disease free remission samples. Disease free remission was defined as minimal residual disease levels of less than $1x10^{-3}$ as measured by the disease specific *IGHV* rearrangements using quantitative PCR or FISH negativity if a marker was available. The analysis of a non-tumor (or germline) sample from the same patient is necessary to correctly identify somatic variants. Currently and for the forseeable future, single nucleotide polymorphism (SNP) data bases will not be comprehensive enough to contain every germline SNP in the population. There are about 1 to 3 million SNPs differing between any two unrelated individuals accounting for about 0.1 % of the genome (Levy *et al.* 2007). The fact that a remission sample was required for proper analysis meant that we could only analyze patients that came into remission after treatment. We thus selected for samples from patients with a more favorable course of disease. Patient characteristics kindly provided by Hofmann-La Roche confirmed this assumption as our exome cohort represents patients with significantly better prognostic markers (*IGHV*, ZAP-70 and Binet stage) compared with the whole study group.

After exome capturing paired-end sequencing was performed on an Illumina Genome Analyzer IIx. We obtained on average 94 million reads per lane and per exome, corresponding to more than 7 Gb of sequence assuming a 80 base pair read length. To align the sequence reads to the genome we chose the Burrows Wheeler Alignment tool, as it is designed to map short reads and allows mismatches and gaps for calling of small insertions and deletions (Li and Durbin 2009). The alignment output was in SAM and BAM format files, which can be manipulated by the SAMtools software package (Li *et al.* 2009). Before starting with variant calling, reads had to pass several quality filters. In a first step, reads derived from clusters on the flow cell with a low signal to noise ratio were discarded (chastity filter). On average, 18.4 % of the reads were discarded. More than 97 % of the remaining reads could be mapped to a unique position in the genome. We further filtered by mapping quality and removed PCR duplicates (11 % on average) since PCR introduced errors may result in false positive variant calls. This meant that we were able to use about 4.8 Gb of sequence data per patient for the downstream analysis. All custom scripts and data analysis workflows were executed on a

Galaxy server, an open source, web-based platform (Goecks *et al.* 2010). For this project we performed no quality trimming, realignment or base quality recalibration.

5.1.2 Downstream Analysis

Downstream analysis was performed using the human RefSeq genes as downloaded in May 2012 from the UCSC table browser based on the human genome assembly from March 2006 NCBI Build. The number of bases on or near target (±250 nucleotides) was about 72.72 %. For all 50, exomes the average on target coverage ranged between 32 and 73 fold. Off note, around 5 % of all targets were not covered at all and 10 % had a coverage of less than 10 fold. Thus, for 85 % of all target positions the coverage was at least 10-fold, sufficient for variant calling (SNVs and InDels). For variant calling, VarScan 2 was used with different settings. We obtained most variants with a minimal variant frequency of 5 % set in the remission sample. The best validation rate of 94 % (true positive variants) was obtained for our somatic workflow.

In total, we could validate 271 variants in the 25 CLL samples including point mutations and small InDels. All variants were located in coding regions except for one splice site mutation affecting the *RB1* gene. We identified, on average, 11 (±4.3) mutations per patient with a minimum of n=3 and a maximum of n=20.

When we take into account that 15 % of all target positions could not be analyzed, as they did not reach the minimum coverage requirements for a SNV to be called, the mutations described are derived from only 85 % of the target (i.e. the RefSeq hg18 genes). This implies that the true number of missense, nonsense and InDel mutations in each sample is about 17 % greater. Considering the range of 3 to 20 mutations per sample, it is likely that between 0 and 3 additional mutations are present in each sample.

This result is in good agreement with published studies on CLL whole exome or whole genome sequencing (Puente *et al.* 2011; Quesada *et al.* 2012). Thus our data indicates that we successfully established whole exome sequencing in our laboratory, starting from library preparation to downstream computational analysis. As we created a solid foundation for whole exome sequencing analysis, we can now further optimize downstream analysis. In some regions for example there can be a great bias between forward and reverse reads. Therefore, only the forward or the reverse reads map to the reference sequence. This so-called strand imbalance is a result of the exome capturing mechanism (Guo *et al.* 2012). Variants that are derived from one strand only are commonly considered to be false positives due to sequencing artifacts, but Guo *et al* showed that strand imbalances, which derive as artifacts from the exome capturing mechanism, have little effect on genotype quality. In our dataset we were able to validate several mutations with the variant reads derived from one strand only, given that there was also a strand imbalance in the reference reads. Therefore variant reads derived from one strand only should not strictly be excluded.

Most false positive calls occurred in the subtraction workflow when the minimum coverage required for the remission sample was set to one. With this setting we called many putative

somatic mutations in the tumor sample that later proved to be germline variants when analyzed with Sanger sequencing (false positives). In our somatic workflow, the minimal coverage requirement for calling SNVs in the remission control was set to ten therefore we faced less such problem to reduce the false positive rate.

In general, a higher coverage will reduce false positive rates. Coverage is dependent on enrichment efficiency. Optimized sample preparation can positively affect coverage of sparsely covered targets. There are many reasons for some target regions to be sparsely covered. One of these reasons can be PCR amplification bias. Aird and colleagues have shown that PCR amplification bias can be reduced by using alternative polymerases and amplification settings (Aird *et al.* 2011). For one of our samples we changed amplification settings, we extended the initial elongation step and the denaturation step at each cycle for the amplification of the adapter ligated libraries. For the amplification of the captured DNA, we added two more PCR cycles (3.5.3.4 and 3.6.1.2). With these settings we could significantly increase the coverage for *NOTCH1*. In the 24 samples, where we did not optimize the PCR settings, the coverage for half of the nucleotide positions in *NOTCH1* was below ten. In contrast, in the one sample where we used optimized PCR-settings, half of the *NOTCH1* positions were covered 30-fold or higher and only 15 % of the positions were covered 10 fold or less. As *NOTCH1* is a frequent mutational target in CLL we also analyzed *NOTCH1* in all 25 samples by Sanger sequencing.

In summary we had a very good validation rate, visual inspection of aligned reads further helped to reduce false positives. A change of downstream analysis settings may reduce the rate of false positive variants but can lead to an increase in the false negative rate, i.e. important somatic mutations will be discarded in the workflow.

5.1.3 Sanger Sequencing as Validation and Screening Tool

Sanger sequencing was used to validate all nonsynonymous mutations independently. Although Sanger sequencing is still the gold standard in diagnostics, its usefulness as an independent validation tool for mutations discovered with NGS approaches has its limitations. Firstly Sanger sequencing has a sensitivity of around 15 %, which means that heterozygous mutations present in clones that constitute less than 30 % cannot be reliably validated by Sanger sequencing. So true positive somatic mutations might be discarded after attempted Sanger sequencing validation. Secondly, Sanger sequencing is time consuming and labor-intensive, taking several days from primer design to final sequence. Thirdly, one loses several nanograms of valuable patient DNA with each Sanger sequencing reaction. For our transcriptome project for example (Greif *et al.* 2011) we screened three candidate genes in 95 patients for mutations. Altogether 46 amplicons per patient sample were amplified for a total of 4370 PCR-products. Sanger sequencing was performed bidirectional for all products resulting in over 8000 sequencing reactions that were performed and analyzed. This was not only laborious and costly, it also required about 1 µg of DNA per patient. Therefore, newer strategies using commercially designed gene panels and small scale NGS approaches are preferable for these focused screening approaches. With these strategies (e.g. the Haloplex

system from Agilent Technologies) several target genes and samples can be sequenced in parallel and with much higher sensitivity.

5.2 Mutations Discovered in the 25 CLL Cases

During the last years the identification of prognostically relevant mutations has not only improved the prediction of disease outcome in CLL but has also remarkable improved our understanding of CLL biology. Although the *IGHV* mutational status and cytogenetic markers are able to predict time to progression, time to therapy and overall survival in large cohorts of CLL patients, the clinical course of the individual patient is very difficult to predict. Furthermore, common FISH markers can be instable and the detection of the *IGHV* mutational status is very laborious and difficult and is not performed routinely in many laboratories. Thus the search for new, relevant biomarkers in CLL is an ongoing endeavor with exome sequencing having become a viable approach to identify somatic mutations in CLL.

5.2.1 Frequency of Nonsynonymous Mutations in CLL

In our study we evaluated 25 previously treated and relapsed CLL patients and discovered a total of 271 point mutations and InDels affecting 260 genes with a mean frequency of eleven mutations per patient. We and others find relatively few mutations per CLL patient in comparison to patients with solid tumors (Quesada *et al.* 2012; Kandoth *et al.* 2013; Vogelstein *et al.* 2013).

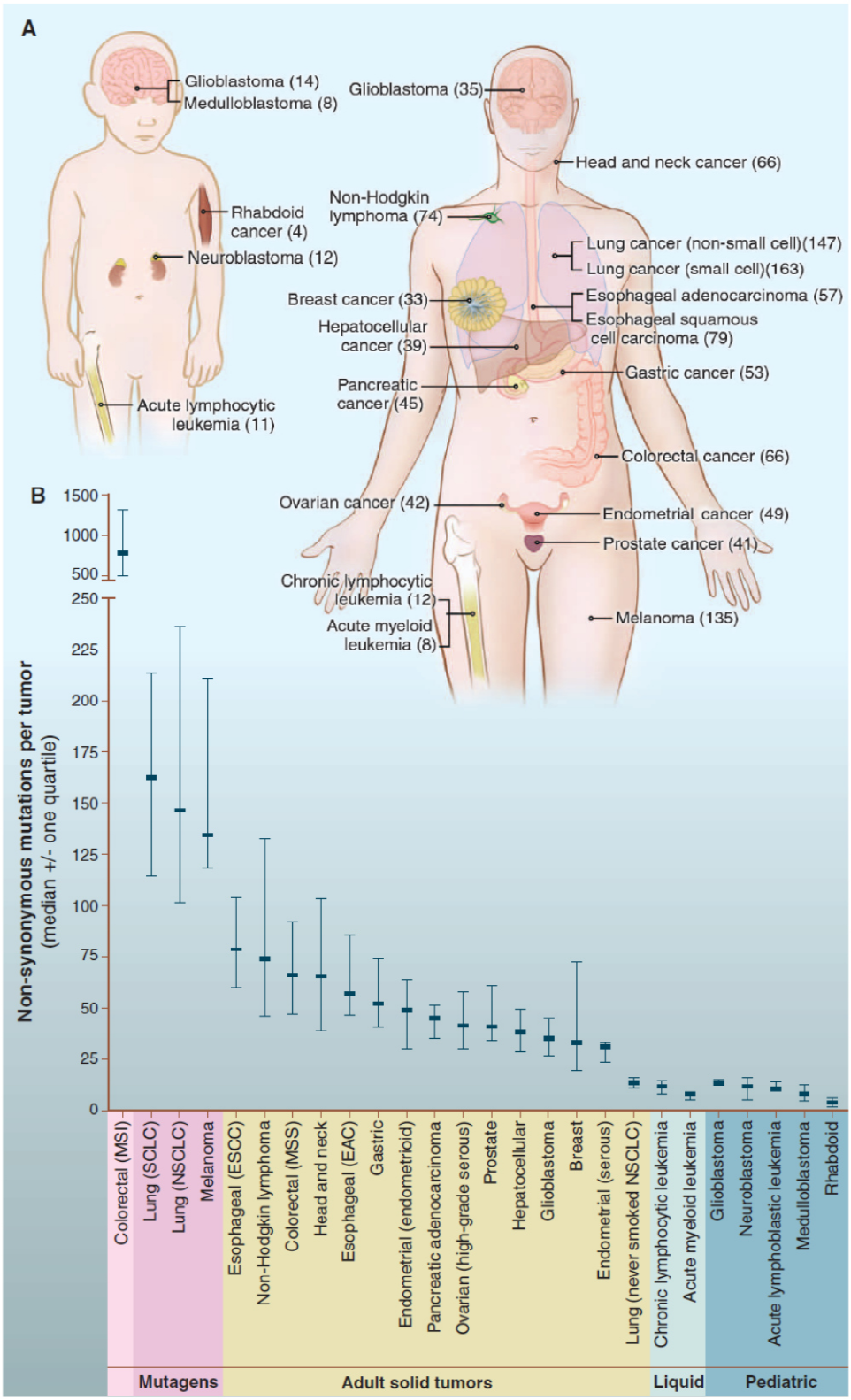


Figure 33 Number of somatic mutations in a variety of cancer types (taken from Vogelstein *et al*, 2013). A) Pediatric cancers (left) and adult cancers (right) with median number of nonsynonymous mutations per tumor in parentheses. A variety of cancer types with median number of nonsynonymous mutations represented as boxplots. MSI: microsatellite instability; SCLC: small cell lung cancers; NSCLC: non-small cell lung cancers; ESCC: esophageal squamous cell carcinomas; MSS: microsatellite stable; EAC esophageal adenocarcinomas.

Although leukemias are a heterogeneous group, AML (Welch and Link 2011; Greif *et al.* 2012) and CLL have a similar, average number of mutations in the coding regions. AML evolves from hematopoietic stem cells, which accumulate mutations that lead to increased proliferation and a block of differentiation. AML is characterized by rapidly proliferating cells and a rapid disease course. In contrast, the cell of origin in CLL is a more differentiated B cell, and disease arises mainly in elderly patients and is characterized by a protracted disease course with the slow accumulation of mature looking B cells (Chiorazzi *et al.* 2005). When we analyzed our first CLL exomes we were surprised that we found such a small number of mutations, especially in the cases with hypermutated *IGHV* genes where we expected more mutations in non-immunoglobulin genes. Indeed, we found that the average number of somatic mutations was slightly higher in the *IGHV* mutated subgroup but the difference was not statistically significant most likely due to the small number of patients we analyzed. Other groups have shown a significant difference in the number of somatic mutations in non-immunoglobulin genes between mutated and unmutated CLL samples (Quesada *et al.* 2012). Landau and colleagues could also show that age is associated with a higher number of somatic mutations (Landau *et al.* 2013).

5.2.2 Recurrently Mutated Genes

If a gene is found mutated more often in cancer in general or in a particular cancer subtype than would be expected by chance, it can be assumed that such a gene harbors driver mutations. Among the 259 mutated genes (mutations in coding regions), we find six of them recurrently mutated in our 25 patients: *XPO1* (16 %), *SF3B1* (12 %), *KRAS* (8 %), *CDH12* (8 %), *PCLO* (8 %), *MYH2* (8 %). Due to our small sample size of only 25 patients and an average of 11 mutated genes per sample, the likelihood that we would find the same gene mutated twice by chance in our cohort should be very low. However, calculating the likelihood of any gene being mutated more than once in a group of 25 patients is not trivial since one has to take into account the mutational target, i.e. the size of the gene that it presents to mutating agents. Thus very large genes like *PCLO* (4935 amino acids) and *MYH2* (1941 amino acids) might be mutated twice in such a cohort by chance without being driver genes.

When we assigned the 259 genes which had been found mutated in coding regions in our WES cohort using the online GSEA tool (http://www.broadinstitute.org/gsea/), we found mutated genes (n=43; 16 %) in all eight categories within this database: cytokines and growth factors, transcription factors, homeodomain proteins, cell differentiation markers, protein kinases, translocated cancer genes, oncogenes and tumor suppressors. A gene can also be assigned to more than one group. Out of the 43 genes, we most frequently see mutated transcription factors (33 %) and protein kinases (28 %). 84 % out of our 25 CLL samples had at least one gene mutated that belonged to one of these 8 categories (none of these genes were found mutated in Patients: P13, P17, P18 and P20).

The nuclear export protein XPO1 was most frequently mutated in our WES cohort. This gene is a putative driver in CLL, which might be more common in relapsed patient samples due to the

preceding treatment. As we found *XPO1* unexpectedly frequently mutated in 41 of 445 (9.2 %) patients, the role of XPO1 in CLL is discussed in a separate chapter.

SF3B1 was found mutated in 3/25 samples (12 %) with various residues affected as shown in Table 14. All three samples had an unmutated *IGHV* status, two had a del(13q), and in one patient none of the FISH markers used showed an abnormality. The first next generation sequencing studies discovered that *SF3B1* is frequently mutated in CLL and also in MDS (Papaemmanuil *et al.* 2011; Wang *et al.* 2011). Since these initial findings, much effort has been put into the evaluation of the role of *SF3B1* mutations as a driver in CLL. SF3B1 is part of the spliceosome, which consists of several subunits of small nuclear ribonucleoprotein complexes, the so called snRNPs (U1, U2, U4, U5 and U6) and numerous splicing factors. The protein SF3B1 is part of one of the two heteromeric protein complexes of the U2 snRNP, namely SF3B, which also contains six other proteins (Wahl *et al.* 2009). Mutations in *SF3B1* are predominantly clustered in the C-terminal region of the protein (Groves *et al.* 1999), which consists of 22 non-identical tandem HEAT repeats (HEAT named after the proteins Huntingtin, elongation factor 3, protein phosphatase 2A and TOR1 kinase), with K700E as the most frequent non-silent mutation. Analyses of diverse CLL cohorts have consistently shown that *SF3B1* mutations are associated with more aggressive disease and poorer clinical outcome (Wang *et al.* 2011; Quesada *et al.* 2012; Oscier *et al.* 2013). Landau and colleagues observed that mutations in *SF3B1* are predominantly subclonal and thus more likely to be involved in disease progression rather than disease initiation (Landau *et al.* 2013). This is consistent with the finding that fludarabine treatment refractory patients have a greater incidence of *SF3B1* mutations (≈17 %) than newly diagnosed patients (Rossi *et al.* 2011). Mutations of the *SF3B1* gene are frequently associated with deletions of the long arm of chromosome 11 (del(11q)). These findings suggest that mutations in *SF3B1* and in *ATM*, which is affected by 11q deletions, might collaborate in CLL pathogenesis (Wang *et al.* 2011). Indeed a siRNA screen identified spliceosome components to be involved in DNA damage response being important for genome stability (Paulsen *et al.* 2009).

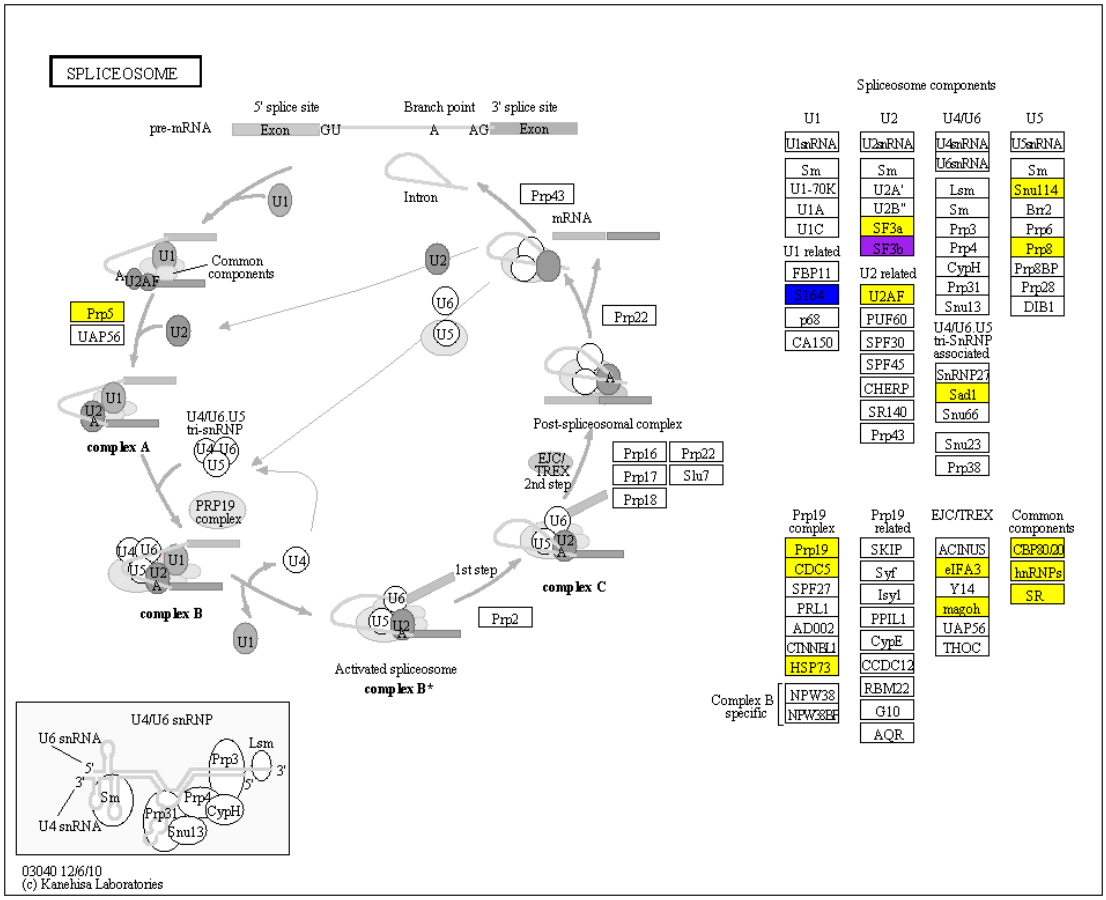


Figure 34 Spliceosome assembly as downloaded from the KEGG database (www.genome.jp/kegg/). Known mutational targets are highlighted in yellow. Blue depicts genes only found in our cohort not in the public CLL datasets (Wang *et al*, Puente *et al* and Quesada *et al*) and purple describes overlapping targets. This map was downloaded from Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

Even before *SF3B1* was discovered to be frequently mutated in CLL and MDS (Papaemmanuil *et al.* 2011; Wang *et al.* 2011) SF3B1 had been shown to be a pharmacological relevant target. Spliceostatin A inhibits *in vitro* splicing and accumulation of pre-mRNA by binding to SF3B1 (Kaida *et al.* 2007). Besides *SF3B1*, many other spliceosome components have been reported to be mutated in hematological malignancies.

We also found *RBM25* (also called S164) to be mutated. Mutations in *RBM25* have not been previously been reported in CLL. Zhou and colleagues showed that the splice factor RBM25 influences apoptosis through the regulation of different *BCLX* isoforms, which act as anti- or pro-apoptotic regulators (Zhou *et al.* 2008). An increase in RBM25 promotes the selection of the pro-apoptotic Bcl-x_S (short form) isoform, and a reduction of RBM25 shifts the balance towards the anti-apoptotic isoform Bcl-x_L (long form). The point mutation p.I94F affects the RRM (RNA recognition motif) domain of the RBM25 protein but it does not seem to be deleterious as phenylalanine and isoleucine both have hydrophobic side chains. The functional prediction tool PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2) predicted the mutation to be benign. Mutations in RBM25 seem to have the potential to be drivers in CLL as it is a regulator of apoptosis.

The members of the *RAS* family of proto-oncogenes, which include *HRAS*, *NRAS* and *KRAS* are frequently mutated in human tumors (≈30 %). RAS proteins are small G-proteins and can switch between a GTP-bound and a GDP-bound state corresponding to the active and inactive state of the protein, respectively. RAS is a key regulator of many signaling pathways including cell growth, differentiation and apoptosis. Most mutations in *RAS* result in the abrogation of the normal GTPase activity resulting in a protein which is constitutively active. In our 25 exome samples we detected one *NRAS* mutation (p.G13V) in a case with unmutated *IGHV* status and Trisomy 12. In addition, two *KRAS* mutations were found: A *KRAS* G13D in a case with unmutated *IGHV* status and a del (11q) and a *KRAS* G12D mutation in a case with mutated *IGHV* status and no FISH marker. Thus, in our exome cohort *RAS* mutations had a frequency of 12 %. *RAS* mutations have been reported to occur only infrequently in CLL (Browett *et al.* 1988; Gougopoulou *et al.* 1996; Domenech *et al.* 2012) compared with other hematological malignancies like AML where they are found with a frequency between 12 % and 27 % (Neubauer *et al.* 2008). As CLL is a very heterogeneous disease, the mutation frequencies for individual genes often vary a lot from study to study. The time point of mutational analysis (first diagnosis vs relapse) and treatment status are variables that have a great impact on the mutation frequency in different cohorts.

We also found recurring mutations in the *PCLO* gene. The PCLO protein is part of the presynaptic cytoskeletal matrix where it is involved in establishing active synaptic zones and vesicle trafficking. The *PCLO* locus has been linked to major depressive disorder (Hek *et al.* 2010). PCLO is a very large protein. Its longest protein coding isoform has 5142 amino acids. We found two non-silent mutations: p.E2041A in a CLL patient with trisomy 12 and mutated *IGHV* status, and p.P2871S in a case with del(13q) and unmutated *IGHV* status. According to the PolyPhen-2 prediction program, the E2041A mutation is probably damaging with a score of 0.975, and the P2871S mutation is possible damaging with a score of 0.775. PCLO was found to be mutated at a frequency of 35 % in a series of 49 large B-cell lymphomas (DLBCL) (Lohr *et al.* 2012). These results make it very likely that mutations in *PCLO* are important for tumorigenesis in lymphoid malignancies.

The human type II classical cadherin-12 (*CDH12*) gene was recurrently mutated in two of our 25 patient samples (N59K and F209L). CDH12 belongs to the cadherin superfamily of calcium dependent transmembrane proteins, which mediate cell-cell adhesion. Cadherins play a role in neural development and have been associated with neurological and psychiatric disorders (Redies *et al.* 2012). Mutations in cadherins have not been linked to leukemias so far. We did not analyze whether this gene is recurrently mutated in our 445 patient extended cohort. Several other cadherin gene family members have been found mutated in other CLL whole exome and genome sequencing studies (Wang *et al.* 2011; Quesada *et al.* 2012)

MYH2, the myosin heavy chain II a gene, was mutated in two patients of our exome cohort. Since *MYH2* presents a large mutational target encoding close to 2000 amino acids, it might be recurrently mutated out of chance. There are, to our knowledge, no reports in the literature that would suggest a functional link between MYH2 mutations and lymphoid malignancies.

Additional studies in larger collectives are warranted to determine the importance of MYH2 mutations as drivers in CLL.

We also extracted all nonsynonymous point mutations and InDels in coding regions from three large CLL genome and exome sequencing projects. 2399 different genes were found to be mutated in the 200 patients analyzed in these three studies (Puente *et al.* 2011; Wang *et al.* 2011; Quesada *et al.* 2012). When we compared the 259 genes mutated (mutations in coding regions) in our 25 CLL samples we found that 82 of these 259 genes (32 %) were also found among the 2399 mutated genes of these three studies. One would have only expected about 10 % of our mutated genes to be also found among these 2399 other genes, if the mutations were distributed randomly across the genome. This result suggests that about 2/3 of these 82 genes are actually driver mutations. Of these 82 genes 47 could only be identified as recurrently mutated when our data was taken into consideration. With this finding, we increase the number of recurrently mutated genes in CLL from 194 to 241. These data emphasize that CLL is an extremely heterogeneous disease at the genetic level with many genes mutated recurrently but at a very low frequency. This is also becoming apparent when we look at the large number of genes (177) that have been found mutated in our small CLL cohort for the first time. For treatment approaches, it is necessary to know the genomic architecture of a disease. Therefore our data significantly increase our knowledge of CLL.

About 241 recurrently mutated genes have been described in CLL up to now. However, only 3 (*SF3B1*, *TP53* and *MYD88*) of them are found mutated in more than 5 % of all the CLL samples of our study and the three public datasets. Other studies also find *NOTCH1 and ATM* frequently mutated (>5 %) in CLL. *SF3B1*, *TP53*, *ATM*, *NOTCH1* and *MYD88* are involved in well-known cellular pathways and core processes in the cell. *Notch* is a transmembrane receptor which gave the name to the intercellular *Notch* signaling cascade, which plays a fundamental role in metazoan development. This signaling cascade controls many cellular processes including proliferation and differentiation. ATM is a serine-threonine kinase, which is involved in cell cycle checkpoint control. ATM is activated upon double strand breaks and phosphorylates proteins that are involved in cell cycle control, apoptosis and DNA repair. One of the target genes of ATM is the tumor suppressor TP53, which preserves genome integrity by regulating growth arrest and apoptosis. MYD88 is a cytosolic adapter protein, which is involved in Toll-like receptor and IL-1 signaling pathway and therefore it is regulating many proinflammatory genes.

Mutations in *SF3B1*, *NOTCH1*, *TP53* and *ATM* are associated with poor prognosis (Austen *et al.* 2005; Zenz *et al.* 2010; Puente *et al.* 2011; Wang *et al.* 2011; Dufour *et al.* 2013). In our samples, the mutational hotspot (p.P2515Rfs*4) in *NOTCH1* was found mutated in only one patient with an unmutated *IGHV* status and no FISH abnormalities. There was also only one patient (P21) with a *TP53* mutation (p.H193R). This patient was normal on FISH and had mutated *IGHV* genes. As our cohort represents mostly patients with a favorable prognosis, this result was not surprising.

5.3 The Relevance of XPO1 in CLL

XPO1, a transport receptor of the karyopherin-β family, was mutated frequently in our CLL cohort (41/445, 9.2 %). XPO1 and other transport proteins share an N-terminal RanGTP-binding motif and interact with NPCs (nuclear core complex). The RanGTPase system plays a major role in nuclear transport, as it ensures the directionality of transport. Exportins like XPO1 bind their cargo in the nucleus, together with RanGTP and the cofactor RanBP3 (activated Ran guanine exchange factor) and move into the cytoplasm where the complex dissociates upon activation of Ran through RanGAP1, the GTPase activating protein, and RanBP1 the coactivator of RanGAP (Sorokin *et al.* 2007). Compared to these, importins (heterodimer of importin-α and importin-β) bind their cargo in the cytoplasm and move them through the NPC into the nucleus where the cargo is released upon binding to RanGTP. The RanGTP/importin complex leaves the nucleus through the NPC without cargo and enters the cytoplasm where RanGTP is removed from the importin (Gorlich and Kutay 1999). The translocation process is most likely based on facilitated diffusion. The directional transport is regulated by Ran.

The human XPO1/CRM1 protein has 1071 amino acid residues and a molecular weight of 112 kDa (Fornerod *et al.* 1997b). This protein is very well characterized and transports a large number of proteins and certain RNA species from the nucleus into the cytoplasm (Fornerod *et al.* 1997a; Stade *et al.* 1997). XPO1 binds to its cargo either directly via Leu rich NESs (nuclear export signals) or indirectly via adapter proteins (Guttler and Gorlich 2011). The entire nuclear export complex has been well characterized by X-ray crystallography (Monecke *et al.* 2013).Till now, more than 200 NES-containing CRM1 cargos have been identified. The CRM1 cargos can be found in a database developed by Xu and colleagues (Xu *et al.* 2012). The consensus sequence of NES was defined by several mutational and computational studies as: ϕ-X_{2-3}-ϕ-X_{2-3}-ϕ-X-ϕ, where ϕ is Leu, Val, Ile, Phe or Met and X can be any amino acid, which is repeated 2-3 times as denoted by the subscript (Bogerd *et al.* 1996; la Cour *et al.* 2004; Dong *et al.* 2009). The tumor suppressor protein TP53 is also among the XPO1 cargos (Freedman and Levine 1998).

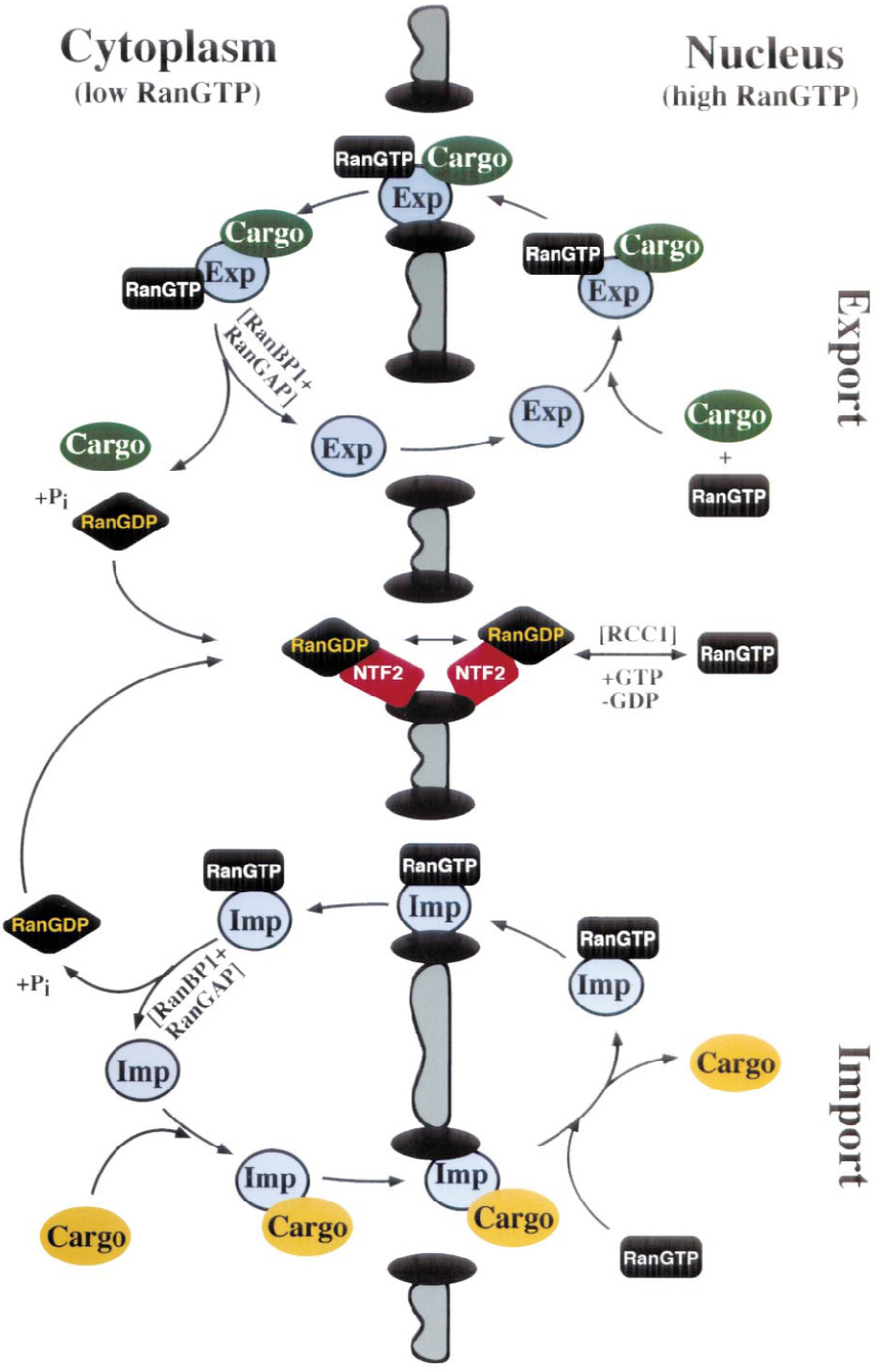


Figure 35 Nuclear Export and Import (taken from Gohrlich and Kutay, 1999). Nuclear export factors (Exp) like XPO1 bind their cargo in the nucleus together with RanGTP and move through nuclear pores into the cytoplasm. When localized in the cytoplasm the low intrinsic GTPase activity of the Ran-Protein is enhanced by RanGAP, which is activated by RanBP1. The complex dissociates, the cargo is now localized in the cytoplasm and the export protein returns into the nucleus. Importins (Imp), in contrast, leave the nucleus while bound to RanGTP. In the cytoplasm RanGAP and RanBP1 accelerate the dissociation of RanGDP from the complex. The free importin can now bind to its cargo and move it into the nucleus. Both transport mechanisms lead to a cytoplasmatic RanGDP enrichment. NTF2, the import receptor of Ran, relocates RanGDP into the nucleus where RCC1, the guanine nucleotide exchange factor, replaces GDP by GTP.

By WES we initially detected three samples (3 of 25, 12%) with *XPO1* mutations, which all affected codon 571. Other studies had detected mutations in codon 571 of *XPO1* as well but with an overall frequency of only 2.4 % (Puente *et al.* 2011). Therefore, we set out to explore this hotspot mutation in a larger group of REACH study patients (n=445) by Sanger sequencing and detected an overall frequency of 9.2 % of *XPO1* mutation (41/445). Of these 41 *XPO1* mutations, 40 affected codon 571. As the screening group also included the 25 exome samples, we detected one additional *XPO1* mutation (E571K) in our exome cohort, which initially went undetected during WES due to the low variant allele frequency of only 7 %. In the screening cohort, we detected a handful of *XPO1* mutations at low variant allele frequency suggesting a later event in CLL evolution. The high frequency of XPO1 mutations in our cohort could be due to the fact that this cohort contains only relapsed and refractory CLLs.

Despite its function as export receptor, XPO1 was initially identified as a protein involved in higher order chromosome structure, which was affected in *XPO1* mutants in fission yeast. Therefore the protein was originally termed CRM1 for chromosome region maintenance protein (Adachi and Yanagida 1989). Subcellular localization studies showed that the protein is not only found in and around the nucleus but also at centrosomes and kinetochors explaining the mitotic phenotype of *CRM1* mutants (Forgues *et al.* 2003; Arnaoutov and Dasso 2005). XPO1 ensures that proteins that are involved in centrosome duplication and kinetochore attachment are correctly positioned. A dysfunction in centrosomes syntheses (centrosome overduplication) and an impaired assembly of microtubules to form the mitotic spindle between kinetochore and the spindle poles may lead to aneuploidy.

5.3.1 Nucleocytoplasmatic Transport and Cancer

The nucleocytoplasmic transport process can be altered in different ways, some of which may promote uncontrolled cell growth, e.g. altered NPCs, modifications in cargo or transport receptors. Alteration of any of these processes may alter the subcellular localization of tumor suppressors, oncoproteins or transcription factors resulting in uncontrolled proliferation. Different fusion proteins have been described in AML that involve nucleoporins, like the translocation t(7:11)(p15;p15.5) resulting in the NUP98-HOXA9 fusion protein or the translocation t(6;9)(p23;q34) leading to a fusion between NUP214, also called CAN, and the nuclear binding protein DEK (Soekarman *et al.* 1992). Many studies focus on the cargo proteins of nucleocytoplasmic transport but not on the nuclear transport machinery itself. However, there is mounting evidence that disruption of the nucleocytoplasmatic transport plays an important role in several cancers. XPO1 was found to be overexpressed in ovarian cancer, cervical cancer, and osteosarcoma. (Noske *et al.* 2008; van der Watt *et al.* 2009; Yao *et al.* 2009). In all three studies, increased XPO1 expression was associated with an unfavorable prognosis. Interestingly, van de Watt and colleagues found that inhibition of XPO1 through Leptomycin B, which covalently binds to the NES-binding groove of XPO1, significantly reduces proliferation of cancer cells but not of normal cells, indicating that certain cancers are dependent on XPO1 to maintain increased proliferation. Several *XPO1* mutations are listed in the COSMIC database (catalogue of somatic mutations in cancer, http://www.sanger.ac.uk). In

CLL, *XPO1* mutations have been identified as driver mutations based on the frequency of XPO1 mutations and the mutational pattern of XPO1 (Landau *et al.* 2013). Interestingly, the highly conserved residue E571 is almost exclusively mutated. Such mutational hotspots are the hallmarks driver mutations that confer an additional function or lead to the formation of dominant negative proteins. The ring-shaped XPO1 protein consists of approximately 20 HEAT repeats and HEAT domains 11 and 12 form the hydrophobic NES-binding groove where residue E571 is located (Dong *et al.* 2009; Sun *et al.* 2013).

5.3.2 Frequency of *XPO1* Mutations in Primary and Relapsed CLL Cases

Mutations in *XPO1*, which is located at chr2p15, are found recurrently in CLL but with a low frequency in the range of 1 to 3 % in cohorts with predominantly *de novo* CLL cases (Puente *et al.* 2011; Wang *et al.* 2011; Balatti *et al.* 2012; Quesada *et al.* 2012). Within our REACH exome cohort (n=25) we initially discovered three *XPO1* mutations. This pointed to a surprisingly high frequency of *XPO1* mutations of 12 %. Therefore we set out to screen this gene in a larger cohort of 445 patient samples. We restricted the screening to the known mutational hotspot in *XPO1*, amino acid E571, as most mutations in XPO1 have been found only at this position. In total, we discovered 41 *XPO1* mutations (9.2 %), 40 mutations at E571 and 1 at E565 (n=33 with E571K, n=4 with E571G, n=2 with E571V, n=1 with E571Q, and n=1 with E565I). Surprisingly, the frequency of 9.2 % *XPO1* hotspot mutations in CLL was much higher than previously described. We may have even underestimated the frequency of *XPO1* mutations as some minor clones might have gone undetected because they were below the detection threshold of our method. A small variant peak in a chromatogram can be either caused by a subclone or be due to a small proportion of tumor material in the sample. The latter possibility could be excluded because we knew the proportion of CD19 positive cells as well as the proportion of FISH marker positive cells (if FISH makers were available in a case). All patients who had a small heterozygous peak in the chromatogram had a high percentage of CD19 positive cells and almost all cells analyzed by FISH were positive if a marker was available. The high percentage of patients with *XPO1* mutations in our cohort, which was not seen in previous reports, might be due to the fact that the other studies analyzed mainly patient samples before administration of treatment. *XPO1* mutations might be associated with disease progression and a more aggressive CLL. All cases with mutated *XPO1* in the other WGS and WES studies had also an unmutated *IGHV* status (Puente *et al.* 2011; Wang *et al.* 2011; Kulis *et al.* 2012). In contrast we found *XPO1* mutations to be significantly associated with 13q deletions, and 13q deletions are associated with a good prognosis.

It should be kept in mind that we looked at a very selected group of patients: CLL patients who had relapsed after treatment. Therefore, the high incidence of *XPO1* mutations in our group could be related to the fact that CLL with *XPO1* mutations are more prone to relapse or that the treatment used in our cohort selected for pre-existing minor clones with *XPO1* mutations. The latter scenario suggests that the frequency of *XPO1* mutations might be greatly underestimated in pre-treatment patients, as the NGS analyses were not sensitive enough to

detect such small clones. As mentioned earlier, the presence of subclonal drivers at diagnosis predict a high probability of relapse in CLL (Landau *et al.* 2013).

5.3.3 Relevance of Mutated *XPO1* in CLL

To our knowledge, there is no data available whether *XPO1* mutations have an impact on the prognosis. We have analyzed overall survival in our *XPO1* screening cohort and did not see a difference between patients with mutations in *XPO1* (including only mutations at codon E571) and those without in the two treatment arms of our study.

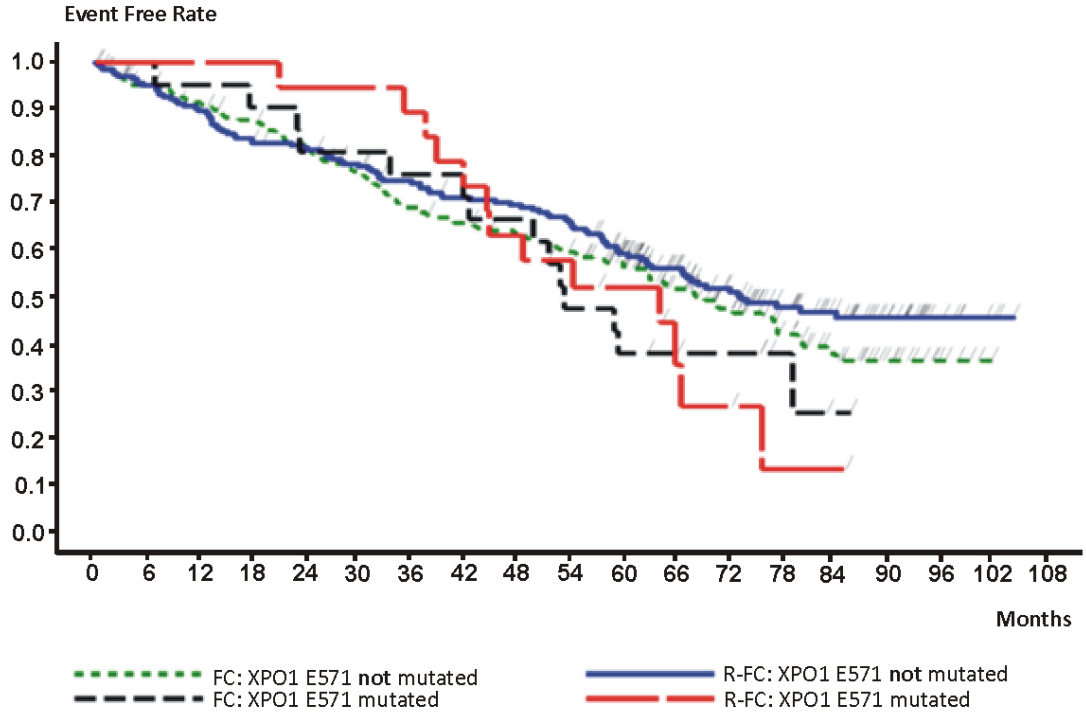


Figure 36 Overall survival of CLL cases with mutated or wild type XPO1 in both treatment arms of the study. Neither in the FC nor in the R-FC arm was a statistically significant difference in overall survival between patients with and without *XPO1* mutations. Prognostic value in FC treated patients: Log-rank test P=0.4259. Data were kindly provided by F. Hoffmann-La Roche.

From these data we can conclude that mutated *XPO1* is not prognostic for overall survival in patients who have not been treated with Rituximab. Furthermore, there was no difference in overall survival for patients with mutated or wild type *XPO1* status when additionally treated with Rituximab. Therefore *XPO1* mutational status was also not predictive for a treatment benefit when adding Rituximab. As *XPO1* might be a marker for disease progression we might not have seen a difference in overall survival within our cohort which represents patients with progressive disease. As mentioned earlier, XPO1 is the main nuclear export factor of the cell, but it is also important for centrosome duplication and attachment of spindle fibers to kinetochores during mitosis. The functional consequences of the mutations at position 571 in XPO1 are not known yet. In most mutated samples the wild type residue at position 571,

namely the negatively charged glutamic acid is replaced by lysine, which carries a positive charge. Since this charge exchange occurs in the cargo binding pocket of XPO1, the binding of cargo proteins might be affected. An impaired separation of chromatides or centrosome overduplication could be the consequence of these mutations and might result in aneuploid tumor cells. During interphase, however, the predominant role of XPO1 as nuclear export factor might be disturbed and therefore many proteins including tumor suppressors and proto-oncogenes might be mislocalized. It has been also shown that XPO1 is involved in U snRNA (small nuclear RNAs with high amount of uracil) export to build up U snRNPs in the cytoplasm. U snRNPs are reimported into the nucleus via snurportin 1. Snurportin 1 itself is recycled to the cytoplasm by XPO1 (Paraskeva *et al.* 1999). However, without experimental data it is highly speculative to predict the functional consequences (gain or loss of function) of mutations affecting residue E571 of XPO1.

In an experiment in *Saccharomyces cerevisiae* with conditional temperature sensitive *xpo1* mutants, generated by random mutagenesis, all mutants were functional at permissive temperature. The mutants showed a prolonged G_1-phase and shortened M-phase. In addition, nuclear export was impaired but this was not restricted to mutations within the NES-binding domain (Neuber 2008). The mutations characterized in this report were not located within the HEAT repeats eleven and twelve. However, also mutations not directly affecting the NES/binding domain may alter cargo binding due to a change in the overall steric arrangement of the NES binding groove. Since XPO1 is involved in the transport of so many proteins, the consequence of impaired XPO1 function is very difficult to predict. As all the mutations found in CLL are heterozygous it can be assumed that a homozygous state of the mutation in *XPO1* would probably be cell lethal.

5.3.4 Therapeutic Targets in Leukemia

For more than a decade, there have been intensive efforts to introduce targeted therapies to accompany and augment cytotoxic chemotherapy. Some of the altered proteins responsible for tumor growth and progression can be influenced by monoclonal antibodies (mABs) or small molecules. The proteins targeted by these compounds can also be present in normal tissue but in cancer cells they are frequently mutated or overexpressed (Gerber 2008). The blocking of tumor growth can be accomplished by several approaches: inhibition of signal transduction, inhibition of proteins regulating the expression of other genes, induction of apoptosis, blocking of angiogenesis, activation of immune response and the delivery of toxic molecules (www.cancer.gov/cancertopics/factsheet/Therapy/targeted).

The ABL tyrosine kinase, which is constitutively activated through the fusion with BCR (BCR/ABL) is a well studied example for the effective use of specific tyrosine kinase inhibitors. By in-vitro screens for tyrosine kinase inhibitors, Imatinib (marketed by Novartis as Gleevec) was found to selectively inhibit the autophosphorylation of BCR/ABL, cKIT and the platelet-derived growth factor receptor (PDGFR). Imatinib blocks tyrosine kinase activity through competition with adenosine triphosphate (ATP) for its binding site. In CML, the treatment of patients with Philadelphia chromosome (>95 %) has been proven to be hugely beneficial

(Druker *et al.* 2001; Kantarjian *et al.* 2002). Treatment with Imatinib can fail when patients have acquired mutations in the *ABL* part of *BCR/ABL* (Jabbour *et al.* 2006), therefore second generation tyrosine kinase inhibitors have been developed that overcome resistance mutations.

Another approach is to target antigens found on the cell surface of cancer cells through monoclonal antibodies. Antibodies may act through the activation of an immune response, interrupting cancer processes or by the delivery of toxins to the target cells. The CD20 surface marker is a tetra-transmembrane protein that is expressed in early B-cells and during development and is lost on plasma cells. CD20 was one of the first therapeutic targets for monoclonal antibodies. The anti CD20 antibody Rituximab was approved 1997. It was manufactured by Genentech/Biogen in Europe and is now also marketed by Hoffmann-LaRoche under the trade name of MabThera. Monotherapy with Rituximab shows therapeutic effects in CLL (Huhn *et al.* 2001) but best results are always obtained together with conventional therapy (Hallek *et al.* 2010). CD20 antibodies are broadly divided into two subgroups: Type I (Rituximab like) mABs cause complement dependent cytotoxicity (non-specific immune system) and redistribute CD20 into lipid rafts; Type II ABs like GA101 (obinutuzumab, manufactured by Hoffman-La Roche) mediate direct cell death. Both antibody types promote antibody-dependent cytotoxicity and phagocytosis (Cragg and Glennie 2004). Fine mapping of the epitopes of Rituximab and GA101 revealed that GA101 binds CD20 in a completely different orientation despite overlapping epitopes (Niederfellner *et al.* 2011). Of course, CD20 monoclonal ABs also bind to normal lymphoid cells and interfere with immune function. Therefore, these antibodies are also used to treat autoimmune diseases.

Also XPO1 is such a potential druggable target, despite the difficulty in predicting the functional consequences of mutated *XPO1*, the overexpression of *XPO1* is found in different tumor entities including AML (Kojima *et al.* 2013) and CLL (Lapalombella *et al.* 2012). The sensitivity of cancer cells to the XPO1 inhibitor Leptomycin B demonstrates the relevance of this pathway for abnormal cell growth. Leptomycin B, originally isolated from a *Streptomyces* species, has contributed to the discovery of hundreds of nuclear export targets. As a therapeutic agent, Leptomycin B proved to be inappropriate due to its strong side effects, which are not mediated by XPO1 inhibition but through unspecific off-target effects (Newlands *et al.* 1996). Leptomycin B covalently binds to cysteine 528 in human XPO1 and contacts 15 additional residues and residue 571 is in contact with the prototypic PKI NES (Sun *et al.* 2013). Novel small-molecules that specifically inhibit XPO1 nuclear export, so called SINEs (selective inhibitors of nuclear export) have recently been developed. Lapalombella *et al* have investigated the efficacy of a SINE called KPT-185 in CLL cells. KPT-185 conjugates to cysteine 528 of XPO1 and showed a cytotoxic effect in primary CLL cells which was stronger than in normal B cells. In addition, enhanced killing was shown in high risk CLL samples (Lapalombella *et al.* 2012).

With next generation sequencing tools we will be able to detect even more druggable targets and will be able to screen large numbers of cancer related genes routinely in a short time.

5.4 NGS as Mutation Discovery and Diagnostic Tool in CLL

CLL is a malignancy that can only be cured by bone marrow transplantation. In many cases, the transformed B cells are very slowly proliferating and treatment is not required in the initial phases of the disease, which can last for decades. As is the case for most malignancies, once a patient relapses after an initial treatment, there is a reduced response to second line treatment. Whole genome and exome sequencing has revealed a number of commonly mutated genes in CLL e.g. *SF3B1* or *NOTCH1*. However, the frequency (≈10 %) of these mutations is relatively low in comparison to the common mutation found in cytogenetically normal AML like *NPM1* mutations (45-64 %) or the *FLT3* internal tandem duplications (28-34 %) (Dohner *et al.* 2010). Thus, we face a so-called long tail distribution with a large number of infrequent driver mutations genes, which mirrors the great heterogeneity of CLL. This situation requires that we analyze even more CLL samples using whole exome or whole genome sequencing (Wood *et al.* 2007; Quesada *et al.* 2013). To avoid a bias in future studies it would be desirable to always secure a germline DNA control sample from each patient at diagnosis, for example a buccal swab. Challenging are also subclonal driver mutations, which have been shown to predict prognosis and clinical outcome (Landau *et al.* 2013). In future studies the presence of potential subclonal drivers should be evaluated at initial diagnosis for risk stratification and to guide treatment strategies.

Certain mutations can also have an impact on the epigenomic landscape as we could demonstrate for secondary AML, where *TET2* and *IDH* mutations correlate with reduced 5-hydroxymethylcytosine levels (Konstandin *et al.* 2011). Whole genome bisulfite sequencing in CLL revealed a DNA methylation signature that distinguished new subtypes (Kulis *et al.* 2012).

As the cost of NGS decreases, it is very likely that whole exome or genome sequencing will become a routine diagnostic tool not only in CLL but in cancer in general. The wide spread use of NGS will reveal hundreds of subgroups in the various pathologically defined tumor entities. Currently we lack the knowledge what the prognosis and treatment response of these many “mutationally” defined subgroups will be. In a genetically heterogeneous disease like CLL, it might be required to associate the plethora of genomic and epigenomic lesions with a few common pathways to obtain larger patient cohorts that can be treated in traditional clinical studies. Eventually, the discovery of ever more small disease subgroups defined by a specific combination of somatic mutation will lead to the development of very personalized treatment strategies for the individual patient.

6 Summary

The advent of next generation sequencing technologies has tremendously improved our understanding of cancer genomics within very short time. Genomic changes on DNA and RNA level can be systematically catalogued, thus the knowledge that we gain will improve our understanding of prognosis, bring us to more detailed diagnostics and increase the number of therapeutic targets.

The aim of this project was to identify new, potentially predictive markers in CLL using whole exome sequencing on an Illumina platform (Illumina, San Diego, CA, USA). Therefore, a cohort of 25 CLL patients and their corresponding disease free remission samples were selected from the CLL REACH trial enclosing 552 previously treated patients. Disease free remission was defined as minimal residual disease levels of less than $1x10^{-3}$ as measured by the disease specific *IGHV* rearrangements using quantitative PCR and FISH negativity if a marker was available. According to flow cytometry data the CLL samples had median of 84 % CD19 positive cells which gives an estimate of the purity of the tumor samples. Flow cytometry measurements, FISH markers and *IGHV* mutational status were routinely determined. The patients selected for our exome cohort had significantly better prognostic markers (*IGHV* mutational status, del(13q), Binet status) than the entire trial, which is based on the fact that we used normal matched control samples from blood at the state of remission. Therefore we lack data from patients with worse prognostic markers who did not reach remission, like patients with a del(17p) or patients with more than one FISH abnormality. Out of the 25 exome patients n=13 harbor a deletion at chromosome 13q, n=5 had no relevant CLL abnormality, n=4 harbored a trisomy 12, n=2 had a deletion at chromosome 11 q and one patient a deletion at locus 6q.

Exome capturing of tumor and non-tumor samples were performed using the Agilent SureSelect target enrichment kit (Agilent Technologies, Santa Clara, CA, USA). Sequencing was performed with 76-80 bp paired-end reads on an Illumina IIx Genome Analyzer (Illumina, San Diego, CA, USA). The mean total number of reads per exome was 94 million (>7Gb per exome), of which in average 72 % could be mapped to the reference genome NCBI36/hg18 with a mapping quality of ≥10. Variant calling and coverage analysis was performed on a reference file containing the human RefSeq genes based on the assembly March 2006 NCBI Build 36/hg18. In total 72.72 % of all bases used for downstream analysis map on or near (±250 bp) target. For all 50 exomes we obtained a mean target coverage of 57 (±11.7) and a median target of 47 (±9.8). More than 90 % of all RefSeq genes were at least covered once. For 85 % of all target positions a coverage of 10-fold and greater was obtained, sufficient enough for variant calling. Target positions covered less than 10-fold or were not covered, which made up around 15 % of the target, could not be used for SNVs or InDel calling in the tumor samples. We called SNVs and InDels specific for the CLL samples by comparing them to their normal, matched control samples from the same patient using VarScan with custom filter settings (Koboldt *et al.* 2009; Koboldt *et al.* 2012). We compared the output of tumor specific non-

synonymous validated SNVs from three different workflows which were called with a p-value of <0.05 and with at least four variant reads from both strands. Using the somatic workflow we obtained best results with n=239 validated nonsynonymous SNVs and n=15 wrong calls for all 25 exomes - that is a positive validation rate of 94 %. Including all validated InDels, nonsynonymous point mutations and splice site mutations, we detected 271 alterations in 260 genes. That is a mean of 11 (±4.3) mutations per patient which is in good agreement with results from other CLL reports. When comparing the number of mutations of U-CLL and M-CLL we did not find a significant difference. In CLL, there is a great diversity of mutated genes among them proto-oncogenes, tumor suppressors, kinases, growth factors, cell differentiation markers and transcription factors possibly inducing leukemogenesis.

In three CLL studies, 2399 different genes were found to be mutated in 200 patients. 194 of these genes were recurrently mutated (Puente *et al.* 2011; Wang *et al.* 2011; Quesada *et al.* 2012). We found an overlap of 82 genes between our CLL cohort and the results from the published datasets. Of these 82 genes, 47 could only be identified as recurrently mutated when our data was taken into consideration. We were thus able to increase the number of known recurrently mutated genes in CLL by more than 20 %.

Frequently mutated genes are always an indication for tumor drivers. Among the recurrently mutated genes, *XPO1* which encodes a nuclear export factor was frequently mutated within our exome cohort with 4 out of 25 samples (16 %). Mutations in this gene almost exclusively affect residue E571. A screening in a larger cohort of patients enclosed in the REACH trial (n=445) revealed a frequency of 9.2 %. Furthermore the *XPO1* mutation was associated with unmutated *IGHV* status and del(13q). Within the REACH *XPO1* screening cohort the mutated gene was neither prognostic for patients treated with FC nor predictive when adding rituximab. Interestingly the frequency in previously treated CLL samples is much higher compared to reports from predominantly primary CLLs. So far the relevance of the *XPO1* mutation affecting amino acid E571 is unclear. As a known druggable target the described mutation may have impact on the effect of XPO1 inhibitors.

In our lab, we have successfully integrated whole exome sequencing technology from sample preparation to bioinformatics analyses starting from matched leukemic and normal blood samples. Exome analysis revealed that CLL is a disease with a broad spectrum of biologically relevant mutational targets. Among all the candidates that could possibly be relevant for leukemogenesis, *XPO1* was the most promising candidate to become a relevant biomarker in CLL. The high frequency of mutations in previously treated patients is indicating that this gene might be involved in disease progression. Further studies will clarify the relevance and function of mutated *XPO1* in CLL.

7 Zusammenfassung

Die Fortschritte in der Hochdurchsatzsequenzierung haben in nur wenigen Jahren unser Verständnis auf dem Gebiet der Krebsgenomforschung signifikant erweitert. Durch die systematische Analyse genomischer Veränderungen von DNA und RNA konnte in kurzer Zeit ein sehr breites Spektrum an Alterationen diverser Tumorgenome dargelegt werden. Diese Daten liefern uns prognostisch relevante Tumormarker sowie therapeutische Targets, was uns in Zukunft eine differenziertere Diagnostik ermöglichen wird.

In dieser Arbeit wurde mittels Exome-Sequenzierung die kodierende Region von 25 CLL Patienten, die schon zu einem früheren Zeitpunkt einmal behandelt wurden, auf Alterationen hin untersucht. Diese Exom-Kohorte, welche aus 25 Patientenproben und den dazugehörigen Keimbahn Kontrollen (Blut zum Zeitpunkt der Remission) besteht, wurde aus einem größeren Kollektiv ausgewählt, welches im Rahmen einer Studie (REACH) behandelt wurde. Patienten der REACH-Studie wurden entweder nur mit Fludarabin und Cyclophosphamid oder zusammen mit dem CD20 Antikörper Rituximab behandelt. Von jedem der 25 Patienten wurde mittels eines kommerziell erwerblichen Kits (Agilent SureSelect target enrichment kit; Agilent Technologies, Santa Clara, CA, USA Agilent Technology) aus genomischer DNA das gesamte Exom zum Zeitpunkt der Erkrankung und der Remission angereichert und anschließend auf einem Illumina IIx Genome Analyzer (Illumina, San Diego, CA, USA) sequenziert. Um zu gewährleisten, dass im Material zum Zeitpunkt der Remission, welches aus Blut gewonnen wurde, möglichst wenige Tumorzellen enthalten sind, wurden nur Proben mit einer minimalen Resterkrankung von weniger als $1x10^{-3}$ CLL-Zellen ausgewählt, die mittels des erkrankungs-spezifischen *IGHV* Rearrangements bestimmt wurde. Außerdem wurden nur Remissionsproben ausgewählt, in denen eine ehemals vorhandene chromosomale Aberration mittels FISH Sonden nicht mehr detektiert werden konnte. Der Anteil an B-Zellen in den Tumorproben wurde bestimmt, indem der Anteil an CD19 positiven Zellen mittels Durchflusszytometrie gemessen wurde, der Median für 23 von 25 gemessenen CLL Proben lag bei 84 %. Die Tatsache, dass wir für die Gesundkontrollen auf peripheres Blut zum Zeitpunkt der Remission zurückgreifen mussten führte in der Exom-Kohorte zu einer signifikanten Anreicherung von Patienten mit günstiger Prognose (mutierter *IGHV* Status, del(13q) und Binet Status B) im Vergleich zur Gesamtkohorte. Unsere Exom-Kohorte besteht aus n=13 Patienten mit del(13q), n=5 zeigten keine für CLL typische chromosomale Aberration, bei n=4 konnte ein Trisomie 12 nachgewiesen werden, in n=2 Proben wurde eine del(11q) festgestellt und bei einem Patient eine Deletion auf Lokus 6q. Patienten mit prognostisch ungünstigeren Markern wie einer del(17p) oder das Vorhandensein mehrerer chromosomaler Aberrationen konnten auf Grund von fehlenden Remissionskontrollen nicht untersucht werden.

Mittels *Paired-End*-Sequenzierung wurden durchschnittlich 94 Millionen *Reads* pro Exom und *Lane* erzeugt, das sind mehr als 7Gb pro Exom wenn man mit einer *Read* Länge von 80 Basen rechnet. Davon konnten wiederum 72 % auf das humane Referenzgenom (NCBI/hg18) aligniert werden, nachdem alle *Reads* mit einer *Mapping*-Qualität kleiner 10 ausgeschlossen wurden.

Die Detektion von Sequenz-Varianten sowie die Abdeckung der einzelnen Ziel-Positionen wurde an Hand einer Referenzsequenz ermittelt, welche alle RefSeq Gene (assembly March 2006 NCBI Build 36/hg18) enthält. Insgesamt konnten 72.72 % der Basen die für die nachgeschalteten Analysen verwendet wurden, auf die RefSeq Gene gemapped werden. Im Mittel wurde eine 57-fache (±11.7) Abdeckung pro Exom erreicht, der Median lag bei 47 (±9.8). Mehr als 90 % der Zielregionen -was den hg18 RefSeq Genen entspricht- wurde mindestens einmal abgedeckt. Für die SNV und InDel Analyse ist eine minimale Abdeckung von 10 *Reads* pro Position erforderlich, was für 85 % der Positionen im Target der Fall war. Die Generierung CLL spezifischer SNVs und InDels erfolgte durch VarScan, einer *Open-Source-Software* (Koboldt *et al.* 2009; Koboldt *et al.* 2012). Mit individuellen Filtereinstellungen wurden drei verschiedene *Workflows* etabliert. Anschließend wurden die nicht-synonymen SNVs mittels Sanger-Sequenzierung validiert und die Ergebnisse der einzelnen *Workflows* miteinander verglichen. Mit dem sogenannten „somatic-workflow“ wurden die besten Ergebnisse erzielt. So konnten für alle 25 CLL Proben 239 nicht-synonyme SNVs bestätigt werden, n=15 wurden nicht bestätigt. Daraus ergibt sich eine Rate von 94 % für richtig bestimmte SNVs. Um die *Workflows* zu vergleichen wurden nur SNVs mit einem P-Wert von <0.05 verglichen, desweiteren mussten die Positionen des varianten Allels von beiden Seiten sequenziert und mindestens durch vier *Reads* abgedeckt sein. Nimmt man alle validierten InDels, nicht-synonyme Punktmutationen und Mutationen an Spleißstellen zusammen konnten wir insgesamt 271 CLL assoziierte Sequenzveränderungen in 260 Genen detektieren. Im Durchschnitt wurden pro Patient 11 (±4.3) Mutationen gefunden, ähnlich wie bei anderen CLL-Kohorten. Im Vergleich von U-CLL und M-CLL, gab keinen signifikanten Unterschied in der Anzahl der Mutationen. In der CLL wurde ein breites Spektrum an mutierten Genen gefunden, darunter einige Protoonkogene, Tumorsuppressoren, Kinasen, Wachstumsfaktoren, Marker für Zelldifferenzierung und Transkriptionsfaktoren. Viele finden sich in bekannten Tumor-Signalwegen, die beispielsweise zu einer unkontrollierten Proliferation oder beeinträchtigter Apoptose führen.

Der Vergleich mit publizierten CLL Genome- und Exomanalysen spiegelt die Heterogenität dieser Erkrankung wider. In drei weiteren CLL-Kohorten wurde das Material von 200 Patienten sequenziert und Mutationen in insgesamt 2399 Genen gefunden, davon wurden Mutationen in 194 Genen bei zwei oder mehr Patienten gefunden (Puente *et al.* 2011; Wang *et al.* 2011; Quesada *et al.* 2012). Stellt man den drei publizierten CLL-Datensätzen die REACH-Kohorte gegenüber, so findet man eine Überschneidung von 82 Genen. Aus diesen 82 Genen konnten 47 identifiziert werden die nur durch unseren Datensatz als rekurrent bezeichnet werden können. Durch den REACH-Exome-Datensatz konnte der Anteil an rekurrent mutierten Gene in der CLL um 20 % erhöht werden.

Häufig mutierte Gene können darauf hinweisen, dass ein Gen für die Tumor/Leukemogenese von Bedeutung ist. Mit vier aus 25 Exom-Proben war *XPO1*, welches einen nukleären Transport-Rezeptor kodiert, das am häufigsten mutierte Gen (16 %). Mutationen in diesem Gen sind fast ausschließlich im Codon E571 zu finden, welches in der hydrophoben NES-Bindestelle liegt. Ein Screening im Bereich des *XPO1* Hotspots in einem größerem Kollektiv der

REACH Kohorte n=445, ergab eine Häufigkeit von 9.2 %. Desweiteren ist die *XPO1* E571 Mutation signifikant mit unmutiertem *IGHV* Status und del(13q) assoziiert. Im Gesamtüberleben der FC-behandelten Patienten zeigte sich jedoch zwischen Patienten mit und ohne Mutation im *XPO1* Gen kein Unterschied, weshalb der Marker für diese Gruppe nicht prognostisch relevant ist. Auch in der Gruppe der zusätzlich mit Rituximab behandelten Patienten konnte für diesen Marker kein Unterschied im Gesamtüberleben festgestellt werden. Daraus ist zu schließen, dass der Marker für eine zusätzliche Behandlung mit Rituximab nicht prädiktiv ist. Es ist zu erwähnen, dass einige der detektierten Mutationen nur in Subklonen zu finden waren. Dies wiederum könnte auch der Grund dafür sein, dass die Rate der beschriebenen *XPO1* Fälle in primären CLL Proben deutlich geringer ausfällt. Auf funktioneller Ebene wurde E571 Mutation noch nicht charakterisiert. Die hydrophobe NES-Bindestelle des Proteins ist auch ein bekanntes Target für diverse Inhibitoren, die für die Krebstherapie relevant sein könnten, auch diesbezüglich ist der Effekt einer Mutation noch nicht untersucht worden.

Mit dieser Arbeit konnten wir zeigen, dass wir die Methode der Exom-Sequenzierung zur Analyse leukämischen Materials und entsprechendem gesunden Kontrollmaterial von der Probenaufarbeitung bis hin zur bioinformatischen Auswertung erfolgreich etabliert haben. Auf Mutationsebene zeigte sich in der CLL ein breites Spektrum an biologisch relevanten Targets. In der CLL könnte zukünftig das häufig mutierte *XPO1* Gen als Biomarker fungieren. Möglicherweise sind *XPO1* Mutationen mit der Progression der Erkrankung assoziiert. Weitere Studien werden die Relevanz und Funktion des mutierten *XPO1* Gens klären.

8 References

Adachi Y, Yanagida M. 1989. Higher order chromosome structure is affected by cold-sensitive mutations in a Schizosaccharomyces pombe gene crm1+ which encodes a 115-kD protein preferentially localized in the nucleus and its periphery. *The Journal of cell biology* 108(4): 1195-1207.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research* 28(20): E87.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biology* 12(2): R18.

Arnaoutov A, Dasso M. 2005. Ran-GTP regulates kinetochore attachment in somatic cells. *Cell Cycle* 4(9): 1161-1165.

Austen B, Powell JE, Alvi A, Edwards I, Hooper L, Starczynski J, Taylor AM, Fegan C, Moss P, Stankovic T. 2005. Mutations in the ATM gene lead to impaired overall and treatment-free survival that is independent of IGVH mutation status in patients with B-CLL. *Blood* 106(9): 3175-3182.

Baccarani M, Dreyling M. 2010. Chronic myeloid leukaemia: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* 21 Suppl 5: v165-167.

Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA *et al.* 2010. Whole exome capture in solution with 3 Gbp of data. *Genome biology* 11(6): R62.

Balatti V, Bottoni A, Palamarchuk A, Alder H, Rassenti LZ, Kipps TJ, Pekarsky Y, Croce CM. 2012. NOTCH1 mutations in CLL associated with trisomy 12. *Blood* 119(2): 329-331.

Barrett JC. 1993. Mechanisms of multistep carcinogenesis and carcinogen risk assessment. *Environmental health perspectives* 100: 9-20.

Binet JL, Auquier A, Dighiero G, Chastang C, Piguet H, Goasguen J, Vaugier G, Potron G, Colona P, Oberling F *et al.* 1981. A new prognostic classification of chronic lymphocytic leukemia derived from a multivariate survival analysis. *Cancer* 48(1): 198-206.

Biomarkers Definition Working Group. 2001. In *Clinical pharmacology and therapeutics*, Vol 69, pp. 89-95. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework.

Bishop JF. 1999. Adult acute myeloid leukaemia: update on treatment. *The Medical journal of Australia* 170(1): 39-43.

Bogerd HP, Fridell RA, Benson RE, Hua J, Cullen BR. 1996. Protein sequence requirements for function of the human T-cell leukemia virus type 1 Rex nuclear export signal delineated by a novel in vivo randomization-selection assay. *Molecular and cellular biology* 16(8): 4207-4214.

Boveri T. 1914. *Zur Frage der Entstehung maligner Tumoren.* Gustav Fisher Verlag, Jena.

Breu H. 2010. A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction., White Paper, Applied Biosystems, Publication 139WP01-02 CO13982.

Browett PJ, Ganeshaguru K, Hoffbrand AV, Norton JD. 1988. Absence of Kirsten-ras oncogene activation in B-cell chronic lymphocytic leukemia. *Leukemia research* 12(1): 25-31.

Chau BN, Wang JY. 2003. Coordinated regulation of life and death by RB. *Nature reviews Cancer* 3(2): 130-138.

Chiorazzi N, Rai KR, Ferrarini M. 2005. Chronic lymphocytic leukemia. *The New England journal of medicine* 352(8): 804-815.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2): 80-92.

Cragg MS, Glennie MJ. 2004. Antibody specificity controls in vivo effector mechanisms of anti-CD20 reagents. *Blood* 103(7): 2738-2743.

Damle RN, Ghiotto F, Valetto A, Albesiano E, Fais F, Yan XJ, Sison CP, Allen SL, Kolitz J, Schulman P *et al.* 2002. B-cell chronic lymphocytic leukemia cells express a surface membrane phenotype of activated, antigen-experienced B lymphocytes. *Blood* 99(11): 4087-4093.

Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, Allen SL, Buchbinder A, Budman D, Dittmar K, Kolitz J *et al.* 1999. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. *Blood* 94(6): 1840-1847.

de Guzman CG, Warren AJ, Zhang Z, Gartland L, Erickson P, Drabkin H, Hiebert SW, Klug CA. 2002. Hematopoietic stem cell expansion and distinct myeloid developmental abnormalities in a murine model of the AML1-ETO translocation. *Molecular and cellular biology* 22(15): 5506-5517.

Dohner H, Estey EH, Amadori S, Appelbaum FR, Buchner T, Burnett AK, Dombret H, Fenaux P, Grimwade D, Larson RA *et al.* 2010. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* 115(3): 453-474.

Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, Bullinger L, Dohner K, Bentz M, Lichter P. 2000. Genomic aberrations and survival in chronic lymphocytic leukemia. *The New England journal of medicine* 343(26): 1910-1916.

Domenech E, Gomez-Lopez G, Gzlez-Pena D, Lopez M, Herreros B, Menezes J, Gomez-Lozano N, Carro A, Grana O, Pisano DG *et al.* 2012. New mutations in chronic lymphocytic leukemia identified by target enrichment and deep sequencing. *PloS one* 7(6): e38158.

Dong X, Biswas A, Suel KE, Jackson LK, Martinez R, Gu H, Chook YM. 2009. Structural basis for leucine-rich nuclear export signal recognition by CRM1. *Nature* 458(7242): 1136-1141.

Druker BJ, Sawyers CL, Kantarjian H, Resta DJ, Reese SF, Ford JM, Capdeville R, Talpaz M. 2001. Activity of a specific inhibitor of the BCR-ABL tyrosine kinase in the blast crisis of chronic myeloid leukemia and acute lymphoblastic leukemia with the Philadelphia chromosome. *The New England journal of medicine* 344(14): 1038-1042.

Dufour A, Palermo G, Zellmeier E, Mellert G, Duchateau-Nguyen G, Schneider S, Benthaus T, Kakadia PM, Spiekermann K, Hiddemann W *et al.* 2013. Inactivation of TP53 correlates with disease progression and low miR-34a expression in previously treated chronic lymphocytic leukemia patients. *Blood* 121(18): 3650-3657.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* 8(3): 175-185.

Fais F, Ghiotto F, Hashimoto S, Sellars B, Valetto A, Allen SL, Schulman P, Vinciguerra VP, Rai K, Rassenti LZ *et al.* 1998. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *The Journal of clinical investigation* 102(8): 1515-1525.

Forgues M, Difilippantonio MJ, Linke SP, Ried T, Nagashima K, Feden J, Valerie K, Fukasawa K, Wang XW. 2003. Involvement of Crm1 in hepatitis B virus X protein-induced aberrant centriole replication and abnormal mitotic spindles. *Molecular and cellular biology* 23(15): 5282-5292.

Fornerod M, Ohno M, Yoshida M, Mattaj IW. 1997a. CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell* 90(6): 1051-1060.

Fornerod M, van Deursen J, van Baal S, Reynolds A, Davis D, Murti KG, Fransen J, Grosveld G. 1997b. The human homologue of yeast CRM1 is in a dynamic subcomplex with CAN/Nup214 and a novel nuclear pore component Nup88. *The EMBO journal* 16(4): 807-816.

Freedman DA, Levine AJ. 1998. Nuclear export is required for degradation of endogenous p53 by MDM2 and human papillomavirus E6. *Molecular and cellular biology* 18(12): 7288-7293.

Gerber DE. 2008. Targeted therapies: a new generation of cancer treatments. *American family physician* 77(3): 311-319.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J *et al.* 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome research* 15(10): 1451-1455.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C *et al.* 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27(2): 182-189.

Goecks J, Nekrutenko A, Taylor J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 11(8): R86.

Gorlich D, Kutay U. 1999. Transport between the cell nucleus and the cytoplasm. *Annual review of cell and developmental biology* 15: 607-660.

Gougopoulou DM, Kiaris H, Ergazaki M, Anagnostopoulos NI, Grigoraki V, Spandidos DA. 1996. Mutations and expression of the ras family genes in leukemias. *Stem Cells* 14(6): 725-729.

Graur D, Li W-H. 2000. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, Mass.

Greif PA, Dufour A, Konstandin NP, Ksienzyk B, Zellmeier E, Tizazu B, Sturm J, Benthaus T, Herold T, Yaghmaie M *et al.* 2012. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood* 120(2): 395-403.

Greif PA, Eck SH, Konstandin NP, Benet-Pages A, Ksienzyk B, Dufour A, Vetter AT, Popp HD, Lorenz-Depiereux B, Meitinger T *et al.* 2011. Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* 25(5): 821-827.

Groves MR, Hanlon N, Turowski P, Hemmings BA, Barford D. 1999. The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* 96(1): 99-110.

Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, Zheng W, Li C. 2012. Exome sequencing generates high quality data in non-target regions. *BMC genomics* 13: 194.

Guttler T, Gorlich D. 2011. Ran-dependent nuclear export mediators: a structural perspective. *The EMBO journal* 30(17): 3457-3474.

Hallek M, Fischer K, Fingerle-Rowson G, Fink AM, Busch R, Mayer J, Hensel M, Hopfinger G, Hess G, von Grunhagen U *et al.* 2010. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet* 376(9747): 1164-1174.

Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK. 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood* 94(6): 1848-1854.

Hanahan D, Weinberg RA. 2000. The hallmarks of cancer. *Cell* 100(1): 57-70.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* 144(5): 646-674.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D *et al.* 2006. GENCODE: producing a reference annotation for ENCODE. *Genome biology* 7 Suppl 1: S4 1-9.

Hek K, Mulder CL, Luijendijk HJ, van Duijn CM, Hofman A, Uitterlinden AG, Tiemeier H. 2010. The PCLO gene and depressive disorders: replication in a population-based study. *Human molecular genetics* 19(4): 731-734.

Huhn D, von Schilling C, Wilhelm M, Ho AD, Hallek M, Kuse R, Knauf W, Riedel U, Hinke A, Srock S *et al.* 2001. Rituximab therapy of patients with B-cell chronic lymphocytic leukemia. *Blood* 98(5): 1326-1331.

Iwasaki H, Akashi K. 2007. Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* 26(6): 726-740.

Jabbour E, Kantarjian H, Jones D, Talpaz M, Bekele N, O'Brien S, Zhou X, Luthra R, Garcia-Manero G, Giles F *et al.* 2006. Frequency and clinical significance of BCR-ABL mutations in patients with chronic myeloid leukemia treated with imatinib mesylate. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* 20(10): 1767-1773.

Kaida D, Motoyoshi H, Tashiro E, Nojima T, Hagiwara M, Ishigami K, Watanabe H, Kitahara T, Yoshida T, Nakajima H *et al.* 2007. Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. *Nature chemical biology* 3(9): 576-583.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA *et al.* 2013. Mutational landscape and significance across 12 major cancer types. *Nature* 502(7471): 333-339.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28(1): 27-30.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* 40(Database issue): D109-114.

Kantarjian H, Sawyers C, Hochhaus A, Guilhot F, Schiffer C, Gambacorti-Passerini C, Niederwieser D, Resta D, Capdeville R, Zoellner U *et al.* 2002. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *The New England journal of medicine* 346(9): 645-652.

Kelly LM, Gilliland DG. 2002. Genetics of myeloid leukemias. *Annual review of genomics and human genetics* 3: 179-198.

Kelly LM, Kutok JL, Williams IR, Boulton CL, Amaral SM, Curley DP, Ley TJ, Gilliland DG. 2002. PML/RARalpha and FLT3-ITD induce an APL-like disease in a mouse model. *Proceedings of the National Academy of Sciences of the United States of America* 99(12): 8283-8288.

Klein U, Dalla-Favera R. 2010. New insights into the pathogenesis of chronic lymphocytic leukemia. *Seminars in cancer biology* 20(6): 377-383.

Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L *et al.* 2001. Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells. *The Journal of experimental medicine* 194(11): 1625-1638.

Knudson AG, Jr. 1986. Genetics of human cancer. *Annual review of genetics* 20: 231-251.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25(17): 2283-2285.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* 22(3): 568-576.

Kojima K, Kornblau SM, Ruvolo V, Dilip A, Duvvuri S, Davis RE, Zhang M, Wang Z, Coombes KR, Zhang N *et al.* 2013. Prognostic impact and targeting of CRM1 in acute myeloid leukemia. *Blood* 121(20): 4166-4174.

Konstandin N, Bultmann S, Szwagierczak A, Dufour A, Ksienzyk B, Schneider F, Herold T, Mulaw M, Kakadia PM, Schneider S *et al.* 2011. Genomic 5-hydroxymethylcytosine levels correlate with TET2 mutations and a distinct global gene expression pattern in secondary acute myeloid leukemia. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* 25(10): 1649-1652.

Krober A, Seiler T, Benner A, Bullinger L, Bruckle E, Lichter P, Dohner H, Stilgenbauer S. 2002. V(H) mutation status, CD38 expression level, genomic aberrations, and survival in chronic lymphocytic leukemia. *Blood* 100(4): 1410-1416.

Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, Martinez-Trillos A, Castellano G, Brun-Heath I, Pinyol M *et al.* 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature genetics* 44(11): 1236-1242.

la Cour T, Kiemer L, Molgaard A, Gupta R, Skriver K, Brunak S. 2004. Analysis and prediction of leucine-rich nuclear export signals. *Protein engineering, design & selection : PEDS* 17(6): 527-536.

Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, Sougnez C, Stewart C, Sivachenko A, Wang L *et al.* 2013. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152(4): 714-726.

Lapalombella R, Sun Q, Williams K, Tangeman L, Jha S, Zhong Y, Goettl V, Mahoney E, Berglund C, Gupta S *et al.* 2012. Selective inhibitors of nuclear export show that CRM1/XPO1 is a target in chronic lymphocytic leukemia. *Blood* 120(23): 4621-4634.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G *et al.* 2007. The diploid genome sequence of an individual human. *PLoS biology* 5(10): e254.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M *et al.* 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456(7218): 66-72.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078-2079.

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. 2011. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12): 1739-1740.

Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, Cruz-Gordillo P, Knoechel B, Asmann YW, Slager SL *et al.* 2012. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 109(10): 3879-3884.

Macdonald D, Richardson H, Raby A. 2003. Practice guidelines on the reporting of smudge cells in the white blood cell differential count. *Archives of pathology & laboratory medicine* 127(1): 105.

Mardis ER. 2008a. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG* 24(3): 133-141.

Mardis ER. 2008b. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics* 9: 387-402.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057): 376-380.

Monecke T, Haselbach D, Voss B, Russek A, Neumann P, Thomson E, Hurt E, Zachariae U, Stark H, Grubmuller H *et al.* 2013. Structural basis for cooperativity of CRM1 export complex formation. *Proceedings of the National Academy of Sciences of the United States of America* 110(3): 960-965.

Morrison SJ, Weissman IL. 1994. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* 1(8): 661-673.

Neubauer A, Maharry K, Mrozek K, Thiede C, Marcucci G, Paschka P, Mayer RJ, Larson RA, Liu ET, Bloomfield CD. 2008. Patients with acute myeloid leukemia and RAS mutations benefit most from postremission high-dose cytarabine: a Cancer and Leukemia Group B study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 26(28): 4603-4609.

Neuber M. 2008. Der Kerntransport-Rezeptor Xpo1 ist beteiligt an der Spindelbildung in der Bäckerhefe Saccharomyces cerevisiae. In *Fachbereich Biologie, Chemie, Pharmazie*. Berlin.

Newlands ES, Rustin GJ, Brampton MH. 1996. Phase I trial of elactocin. *British journal of cancer* 74(4): 648-649.

Niederfellner G, Lammens A, Mundigl O, Georges GJ, Schaefer W, Schwaiger M, Franke A, Wiechmann K, Jenewein S, Slootstra JW *et al.* 2011. Epitope characterization and crystal structure of GA101 provide insights into the molecular basis for type I/II distinction of CD20 antibodies. *Blood* 118(2): 358-367.

NIH consensus conference. Treatment of early-stage breast cancer. 1991. *JAMA : the journal of the American Medical Association* 265(3): 391-395.

Noske A, Weichert W, Niesporek S, Roske A, Buckendahl AC, Koch I, Sehouli J, Dietel M, Denkert C. 2008. Expression of the nuclear export protein chromosomal region maintenance/exportin 1/Xpo1 is a prognostic factor in human ovarian cancer. *Cancer* 112(8): 1733-1743.

Nowell PC. 1976. The clonal evolution of tumor cell populations. *Science* 194(4260): 23-28.

Oldenhuis CN, Oosting SF, Gietema JA, de Vries EG. 2008. Prognostic versus predictive value of biomarkers in oncology. *Eur J Cancer* 44(7): 946-953.

Oscier DG, Rose-Zerilli MJ, Winkelmann N, Gonzalez de Castro D, Gomez B, Forster J, Parker H, Parker A, Gardiner A, Collins A *et al.* 2013. The clinical significance of NOTCH1 and SF3B1 mutations in the UK LRF CLL4 trial. *Blood* 121(3): 468-475.

Oscier DG, Thompsett A, Zhu D, Stevenson FK. 1997. Differential rates of somatic hypermutation in V(H) genes among subsets of chronic lymphocytic leukemia defined by chromosomal abnormalities. *Blood* 89(11): 4153-4160.

Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C *et al.* 2011. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *The New England journal of medicine* 365(15): 1384-1395.

Paraskeva E, Izaurralde E, Bischoff FR, Huber J, Kutay U, Hartmann E, Luhrmann R, Gorlich D. 1999. CRM1-mediated recycling of snurportin 1 to the cytoplasm. *The Journal of cell biology* 145(2): 255-264.

Paulsen RD, Soni DV, Wollman R, Hahn AT, Yee MC, Guan A, Hesley JA, Miller SC, Cromwell EF, Solow-Cordero DE *et al.* 2009. A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. *Molecular cell* 35(2): 228-239.

Pollock JL, Westervelt P, Kurichety AK, Pelicci PG, Grisolano JL, Ley TJ. 1999. A bcr-3 isoform of RARalpha-PML potentiates the development of PML-RARalpha-driven acute promyelocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America* 96(26): 15103-15108.

Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M *et al.* 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475(7354): 101-105.

Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, Bassaganyas L, Ramsay AJ, Bea S, Pinyol M, Martinez-Trillos A *et al.* 2012. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature genetics* 44(1): 47-52.

Quesada V, Ramsay AJ, Rodriguez D, Puente XS, Campo E, Lopez-Otin C. 2013. The genomic landscape of chronic lymphocytic leukemia: clinical implications. *BMC medicine* 11: 124.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.

Rai KR, Sawitsky A, Cronkite EP, Chanana AD, Levy RN, Pasternack BS. 1975. Clinical staging of chronic lymphocytic leukemia. *Blood* 46(2): 219-234.

Redies C, Hertel N, Hubner CA. 2012. Cadherins and neuropsychiatric disorders. *Brain research* 1470: 130-144.

Renan MJ. 1993. How many mutations are required for tumorigenesis? Implications from human cancer data. *Molecular carcinogenesis* 7(3): 139-146.

Robak T, Dmoszynska A, Solal-Celigny P, Warzocha K, Loscertales J, Catalano J, Afanasiev BV, Larratt L, Geisler CH, Montillo M *et al.* 2010. Rituximab plus fludarabine and cyclophosphamide prolongs progression-free survival compared with fludarabine and cyclophosphamide alone in previously treated chronic lymphocytic leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 28(10): 1756-1765.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* 29(1): 24-26.

Rossi D, Bruscaggin A, Spina V, Rasi S, Khiabanian H, Messina M, Fangazio M, Vaisitti T, Monti S, Chiaretti S *et al.* 2011. Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: association with progression and fludarabine-refractoriness. *Blood* 118(26): 6904-6908.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365-386.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74(12): 5463-5467.

Schadt EE, Turner S, Kasarskis A. 2010. A window into third-generation sequencing. *Human molecular genetics* 19(R2): R227-240.

Schroeder HW, Jr., Dighiero G. 1994. The pathogenesis of chronic lymphocytic leukemia: analysis of the antibody repertoire. *Immunology today* 15(6): 288-294.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat Biotechnol* 26(10): 1135-1145.

Smith A, Roman E, Howell D, Jones R, Patmore R, Jack A. 2010. The Haematological Malignancy Research Network (HMRN): a new information strategy for population based epidemiology and health service research. *British journal of haematology* 148(5): 739-753.

Soekarman D, von Lindern M, Daenen S, de Jong B, Fonatsch C, Heinze B, Bartram C, Hagemeijer A, Grosveld G. 1992. The translocation (6;9) (p23;q34) shows consistent rearrangement of two genes and defines a myeloproliferative disorder with specific clinical features. *Blood* 79(11): 2990-2997.

Sorokin AV, Kim ER, Ovchinnikov LP. 2007. Nucleocytoplasmic transport of proteins. *Biochemistry Biokhimiia* 72(13): 1439-1457.

Stade K, Ford CS, Guthrie C, Weis K. 1997. Exportin 1 (Crm1p) is an essential nuclear export factor. *Cell* 90(6): 1041-1050.

Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C *et al.* 2011. Comparison of solution-based exome capture methods for next generation sequencing. *Genome biology* 12(9): R94.

Sun Q, Carrasco YP, Hu Y, Guo X, Mirzaei H, Macmillan J, Chook YM. 2013. Nuclear export inhibition through covalent conjugation and hydrolysis of Leptomycin B by CRM1. *Proceedings of the National Academy of Sciences of the United States of America* 110(4): 1303-1308.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*.

Till JE, Mc CE. 1961. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation research* 14: 213-222.

van der Watt PJ, Maske CP, Hendricks DT, Parker MI, Denny L, Govender D, Birrer MJ, Leaner VD. 2009. The Karyopherin proteins, Crm1 and Karyopherin beta1, are overexpressed in cervical cancer and are critical for cancer cell survival and proliferation. *International journal of cancer Journal international du cancer* 124(8): 1829-1840.

Vogelstein B, Kinzler KW. 1993. The multistep nature of cancer. *Trends in genetics : TIG* 9(4): 138-141.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. 2013. Cancer genome landscapes. *Science* 339(6127): 1546-1558.

Wahl MC, Will CL, Luhrmann R. 2009. The spliceosome: design principles of a dynamic RNP machine. *Cell* 136(4): 701-718.

Wang CL, Harper RA, Wabl M. 2004. Genome-wide somatic hypermutation. *Proceedings of the National Academy of Sciences of the United States of America* 101(19): 7352-7356.

Wang L, Lawrence MS, Wan Y, Stojanov P, Sougnez C, Stevenson K, Werner L, Sivachenko A, DeLuca DS, Zhang L *et al.* 2011. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England journal of medicine* 365(26): 2497-2506.

Weissman IL. 2000. Stem cells: units of development, units of regeneration, and units in evolution. *Cell* 100(1): 157-168.

Welch JS, Link DC. 2011. Genomics of AML: clinical applications of next-generation sequencing. *Hematology / the Education Program of the American Society of Hematology American Society of Hematology Education Program* 2011: 30-35.

WHO International Programme on Chemical Safety. 2001. Biomarkers in Risk Assessment: Validity and Validation., Retrieved from http://www.inchem.org/documents/ehc/ehc/ehc222.htm.

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J *et al.* 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318(5853): 1108-1113.

Xu D, Grishin NV, Chook YM. 2012. NESdb: a database of NES-containing CRM1 cargoes. *Molecular biology of the cell* 23(18): 3673-3676.

Yao Y, Dong Y, Lin F, Zhao H, Shen Z, Chen P, Sun YJ, Tang LN, Zheng SE. 2009. The expression of CRM1 is associated with prognosis in human osteosarcoma. *Oncology reports* 21(1): 229-235.

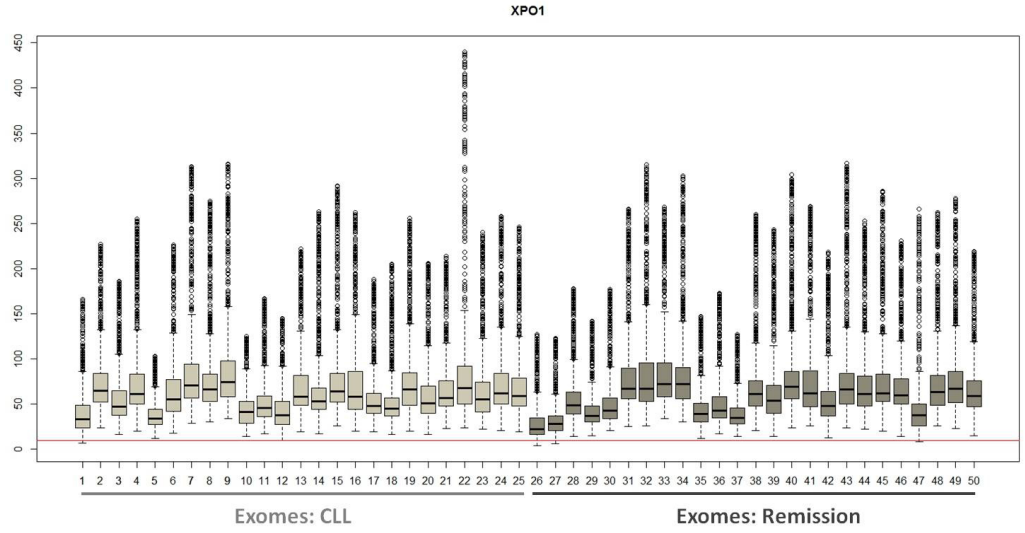
Zenz T, Eichhorst B, Busch R, Denzel T, Habe S, Winkler D, Buhler A, Edelmann J, Bergmann M, Hopfinger G *et al.* 2010. TP53 mutation and survival in chronic lymphocytic leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 28(29): 4473-4479.

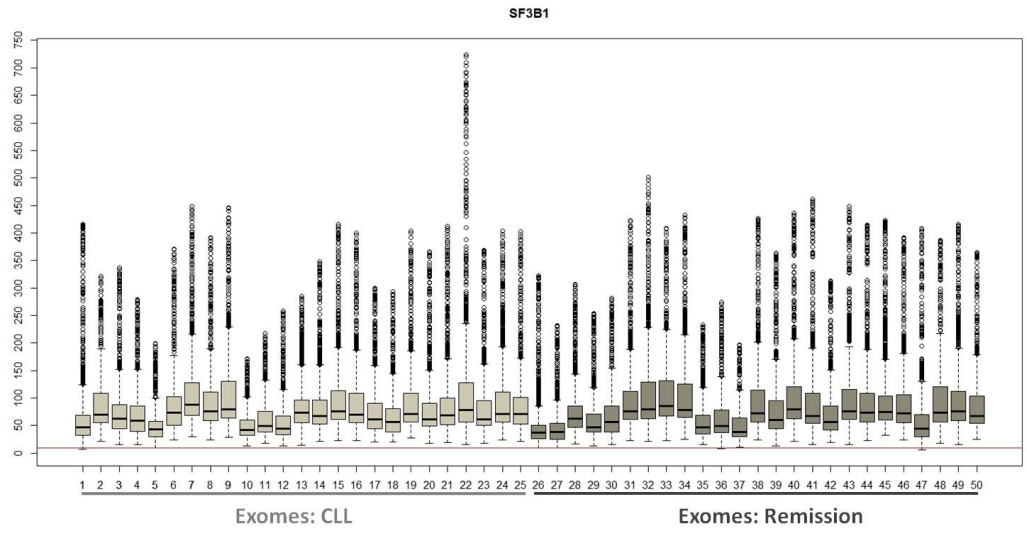
Zhou A, Ou AC, Cho A, Benz EJ, Jr., Huang SC. 2008. Novel splicing factor RBM25 modulates Bcl-x pre-mRNA 5' splice site selection. *Molecular and cellular biology* 28(19): 5924-5936.

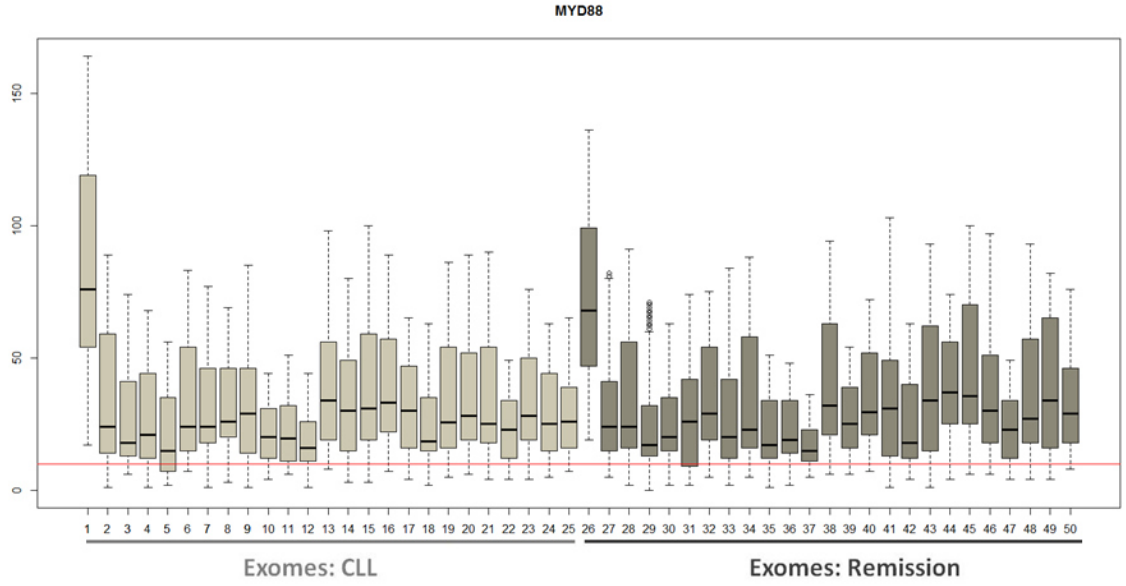
9 Appendix

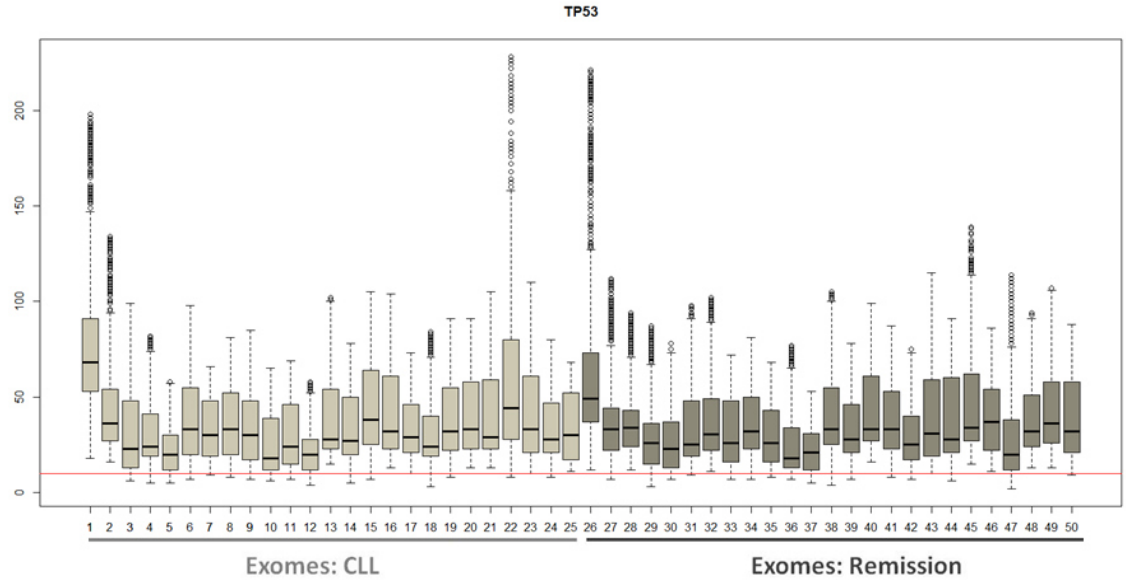
9.1 Per Base Coverage of Frequently Mutated Genes in CLL

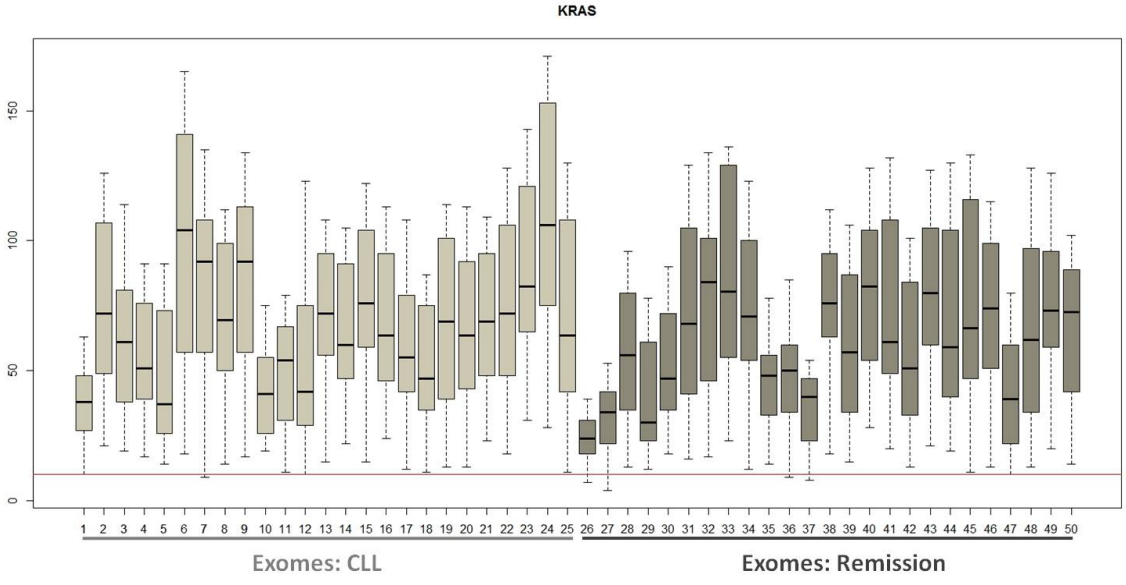
Per base coverage of each exome for frequently mutated genes in CLL and/or recurrently mutated genes as found in the REACH exome cohort. Per base coverage was computed with BEDtools and boxplots generated with R. On y-axis the coverage is depicted and the red line is drawn at ten-fold coverage. The coverage of a gene was calculated for each exome separately, on the left side the CLL samples are represented and on right sight the corresponding remission samples. Exomes 1-25 correspond with CLL patient sample P1-P25 and exomes 26-50 correspond with remission samples P1-P25.

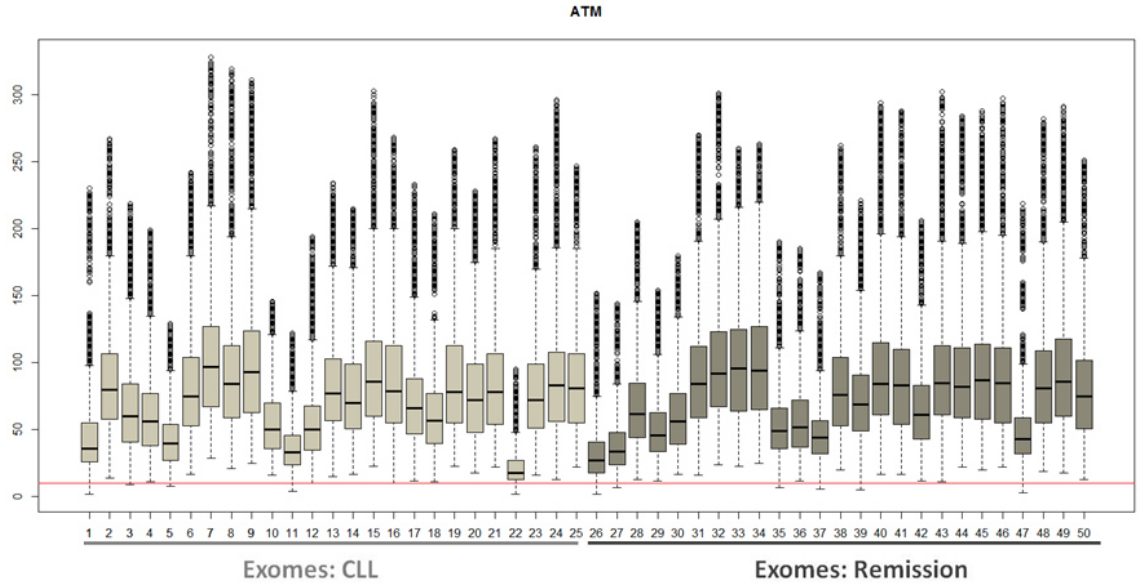


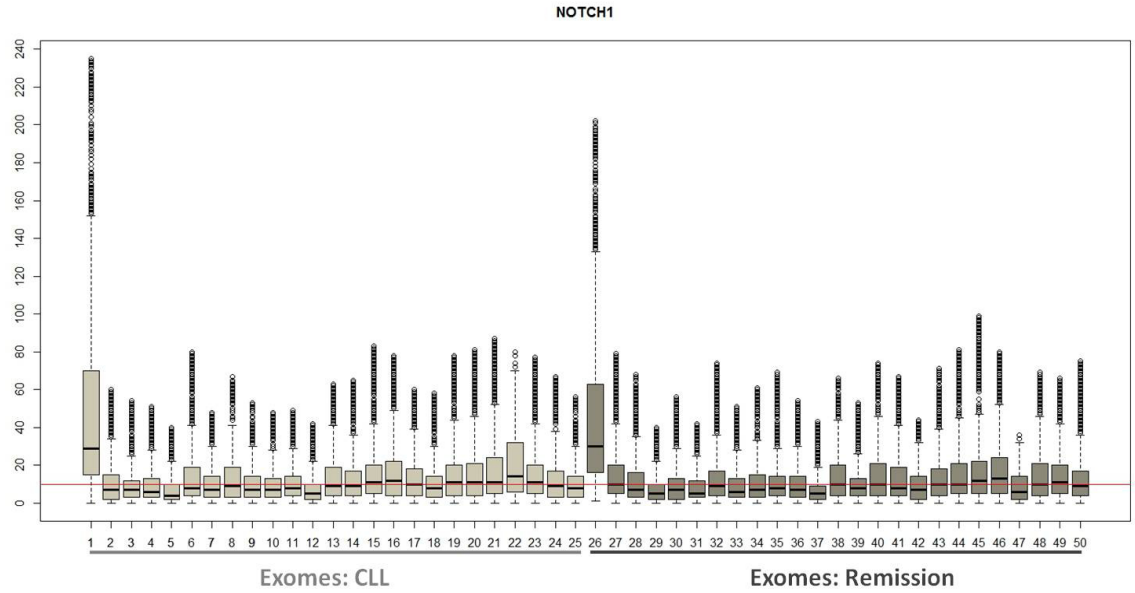


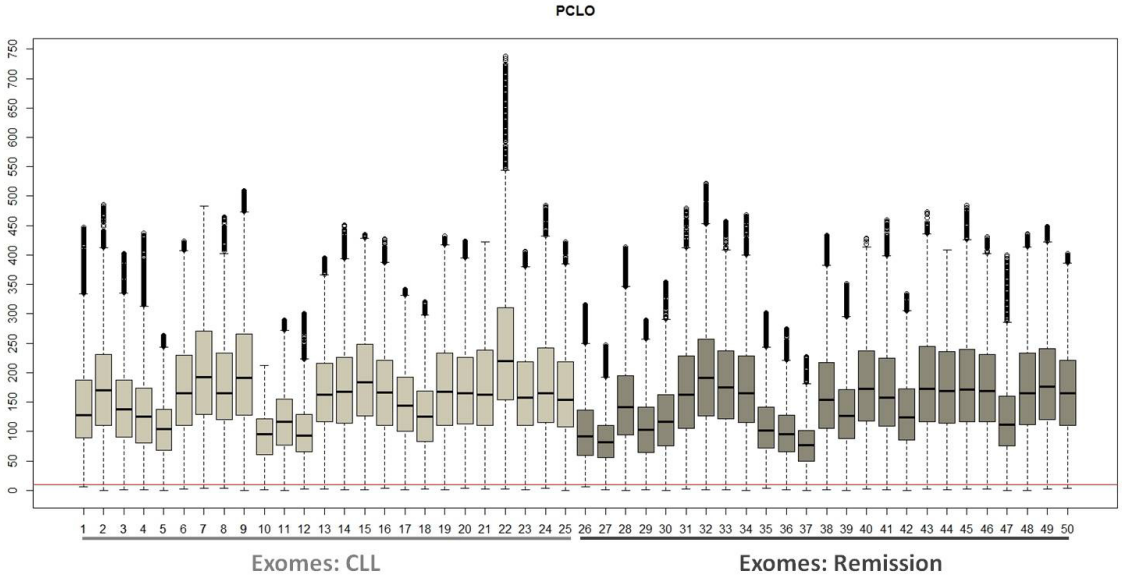

MYD88
0
50
100
150
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission

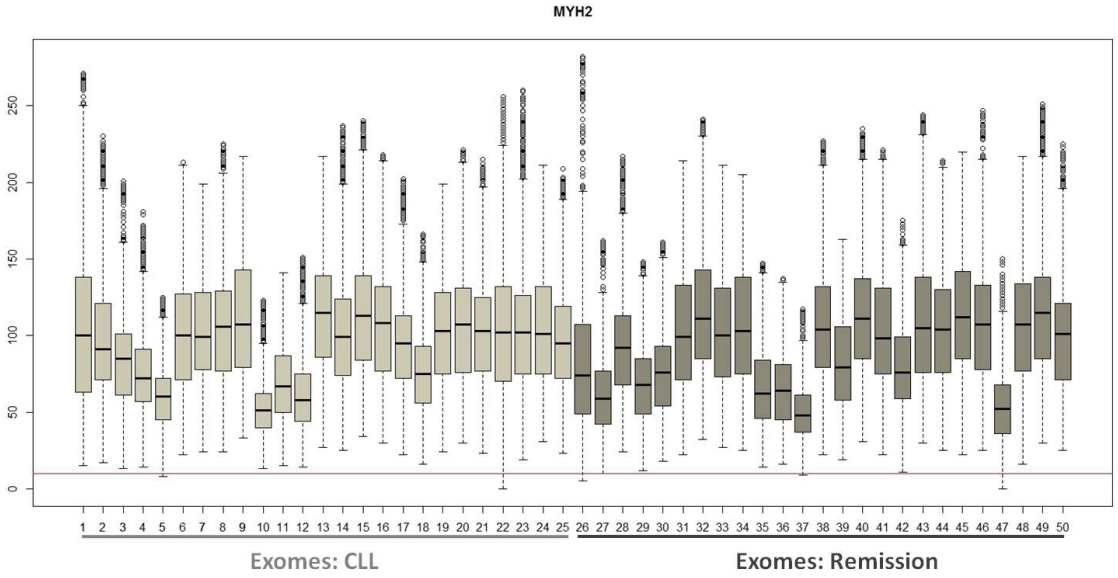

TP53
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission


KRAS
0
50
100
150
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission


ATM
0
50
100
150
200
250
300
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission

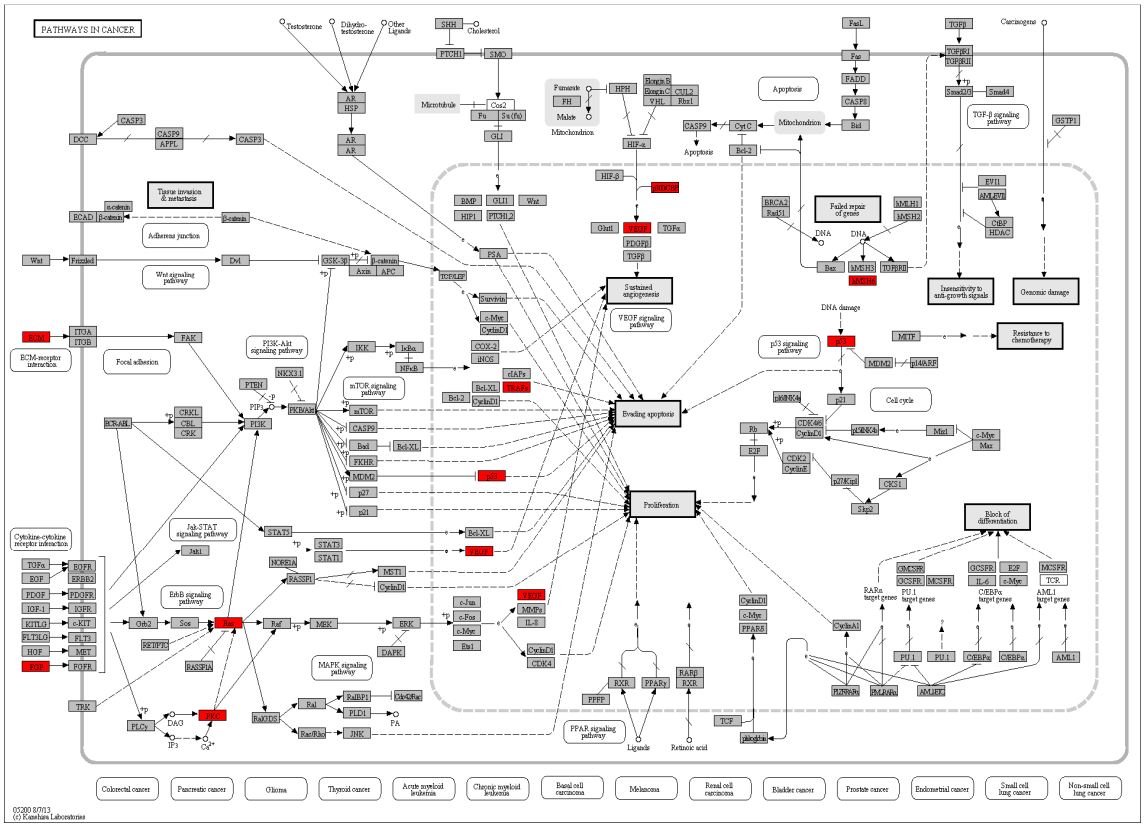

NOTCH1
0
20
40
60
80
100
120
140
160
180
200
220
240
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission


PCLO
0 50 100 150 200 250 300 350 400 450 500 550 600 650 700 750
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission


MYH2
0 50 100 150 200 250
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Exomes: CLL
Exomes: Remission

9.2 Mutations in Components of Cancer Pathways

Mutations found within the 25 exomes in components of cancer pathways: Cancer pathways according to KEGG (http://www.genome.jp/kegg/kegg2.html), including somatically mutated genes in CLL as highlighted in red.



9.3 Demographics of *XPO1* Screening Cohort

Data were kindly provided by F. Hoffmann-La Roche.

Randomized Treatment: FC

	Mutated N = 21	No Mutation N = 196
Sex		
MALE	15 (71%)	130 (66%)
FEMALE	6 (29%)	66 (34%)
n	21	196
Race		
CAUCASIAN	21 (100%)	193 (98%)
BLACK	-	-
ORIENTAL	-	-
OTHER	-	3 (2%)
n	21	196
Age (years)		
Mean	59.6	61.5
SD	8.95	9.04
Median	62.0	62.0
Min-Max	35 - 73	37 - 81
n	21	196
Age (years) Categories		
< 65	14 (67%)	113 (58%)
>=65 - <=70	5 (24%)	51 (26%)
> 70	2 (10%)	32 (16%)
n	21	196
Weight (kg)		
Mean	77.4	76.8
SD	16.12	14.10
Median	76.0	75.0
Min-Max	47 - 119	52 - 126
n	21	196
Height (cm)		
Mean	170.5	169.6
SD	7.71	8.97
Median	171.0	170.0
Min-Max	159 - 184	149 - 188
n	21	193

Randomized Treatment: FC-R

	Mutated N = 19	No Mutation N = 208
Sex		
MALE	10 (53%)	143 (69%)
FEMALE	9 (47%)	65 (31%)
n	19	208
Race		
CAUCASIAN	19 (100%)	204 (98%)
BLACK	-	-
ORIENTAL	-	1 (<1%)
OTHER	-	3 (1%)
n	19	208
Age (years)		
Mean	58.2	62.5
SD	8.85	8.69
Median	59.0	63.0
Min-Max	39 - 69	35 - 80
n	19	208
Age (years) Categories		
< 65	12 (63%)	116 (56%)
>=65 - <=70	7 (37%)	55 (26%)
> 70	-	37 (18%)
n	19	208
Weight (kg)		
Mean	74.2	76.7
SD	11.85	15.23
Median	73.0	76.0
Min-Max	54 - 93	46 - 127
n	19	208
Height (cm)		
Mean	168.6	170.1
SD	11.37	9.79
Median	168.0	170.0
Min-Max	154 - 190	145 - 197
n	19	205

9.4 Disease Characteristics of *XPO1* Screening Cohort

ECOG is a scoring system to quantify cancer patients by their general well-being and activities of daily life. Data were kindly provided by F. Hoffmann-La Roche.

Randomized Treatment: FC

	Mutated N = 21	No Mutation N = 196
Binet Stage		
A	-	19 (10%)
B	16 (76%)	112 (57%)
C	5 (24%)	65 (33%)
n	21	196
B-Symptoms		
YES	8 (38%)	63 (32%)
NO	13 (62%)	133 (68%)
n	21	196
ECOG Status		
0	12 (57%)	111 (57%)
1	9 (43%)	85 (43%)
n	21	196

Randomized Treatment: FC-R

	Mutated N = 19	No Mutation N = 208
Binet Stage		
A	2 (11%)	19 (9%)
B	14 (74%)	122 (59%)
C	3 (16%)	67 (32%)
n	19	208
B-Symptoms		
YES	4 (21%)	58 (28%)
NO	15 (79%)	150 (72%)
n	19	208
ECOG Status		
0	14 (74%)	124 (60%)
1	5 (26%)	84 (40%)
n	19	208

9.5 Abbreviations

°C	degree Celsius
µ	micro ($1x10^{-6}$)
AA	Amino Acid
ALL	Acute Lymphocytic Leukemia
AML	Acute Myeloid Leukemia
APS	Adenosine 5' phosphosulfate
Array CGH	Array-based comparative genomic hybridization
ATP	Adenosin triphosphate
BAM	binary version of SAM file
BAQ	per-Base Alignment Quality
BED	Browser Extensible Data (tabular file format)
bp	base pair(s)
BWA	Burrwos-Wheeler Alignment tool
cDNA	complementary DNA
CIGAR	compact idiosyncratic gapped alignment report
CLL	Chronic Lymphocytic Leukemia
CLP	common lymphoid progenitors
CMP	common myeloid progenitors
CML	Chronic Myeloid Leukemia
dATP	desoxyadenosine triphosphate
dbSNP130	SNP database build 130
ddH_2O	double distilled water
ddNTPs	dideoxynucleotide triphosphates
del	Deletion
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
dsDNA	double stranded DNA
ECOG	Eastern Cooperative Oncology Group
e.g.	Example
EDTA	ethylenediaminetetraacetic acid
FAB	French-American- British classification system for acute leukemia
FC	Fludarabine Cyclophosphamide (treatment regime REACH study)
FC-R	Fludarabine Cyclophosphamide Rituximab (treatment regime REACH study)
FISH	fluoreszenz *in situ* hybridization
g	Gram
Gb	Gigabases
GMP	granulocyte-monocyte progenitor
HEAT	repeat domains in proteins, named after the proteins Huntingtin, Elongation factor 3, protein phosphatase 2A and TOR1 kinase where it was found first
hg18	human genome 18
HS	High Stringency settings for variant calling
HSC	Hematopoietic Stem Cell

HGNC	HUGO Gene Nomenclature Committee
HUGO	Human Genome Organization
i.e.	example
IGHV	immunoglobulin heavy chain variable region genes
IGV	Integrative Genomics Viewer
InDel	to summarize Insertions Deletions
iqrpos	in quantitative range positive
kb	Kilo base pairs
kDa	kilodalton(s)
KEGG	Kyoto Encyclopedia of Genes and Genomes
kHz	Kilo Hertz
l	Liter
LOH	loss of heterozygosity
LS	Low Stringency settings for variant calling
LT-HSC	long-term hematopoietic stem cell
M	Molar
m-	milli ($1x10^{-3}$)
mAB	monoclonal antibody
Mb	mega base pairs
M-CLL	mutated CLL
MEP	megakaryocytic-erythrocyte progenitors
min	minute(s)
miRNA	microRNA
MPP	multipotent progenitors
MQ	mapping quality
MRD	minimal residual disease
mRNA	messenger RNA
MSigDB	Molecular Signatures Database (collection of annotated gene sets)
n	nano ($1x10^{-9}$)
NCBI	National Center of Biotechnology Information
NES	nucleic export signal
NGS	next generation sequencing
NIH	National Institute of Health
NLS	nuclear localization signal
oqrpos	out of quantitative range positive
OS	Overall Survival
p	pico ($1x10^{-12}$)
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PEG	polyethylenglycol
PFS	Progression Free Survival
PPi	pyrophosphate
PhiX	Control library generated from PhiX virus

RNA	ribonucleic acid
rpm	revolutions per minute
SAM	Sequence Alignment/Map format
sd	standard deviation
SINE	Selective inhibitor of nuclear export
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SOLiD	Sequencing by Oligo Ligation Detection
ST-HSC	short-term hematopoietic stem cell
Taq	*Thermus aquaticus*
TBE	Tris-borate-EDTA
TE	Tris-EDTA buffer
Tm	melting temperature
U-CLL	unmutated CLL
UCSC	University of California Santa Cruz (provides Genome Browser)
UV	ultraviolet
WHO	World Health Organization
ZAP-70	zeta-chain associated protein kinase 70

Single letter codes for amino acids

nonpolar

A	**(Ala)**	**Alanine**
V	**(Val)**	**Valine**
L	**(Leu)**	**Leucine**
I	**(Ile)**	**Isoleucine**
M	**(Met)**	**Methionine**
P	**(Pro)**	**Proline**
G	**(Gly)**	**Glycine**

polar hydrophilic

S	**(Ser)**	**Serine**
T	**(Thr)**	**Threonine**
C	**(Cys)**	**Cysteine**
N	**(Asn)**	**Asparagine**
E	**(Gln)**	**Glutamic acid**

aromatic

F	**(Phe)**	**Phenylalanine**
Y	**(Tyr)**	**Tyrosine**
W	**(Trp)**	**Tryptophan**

positively charged

K	**(Lys)**	**Lysine**
R	**(Arg)**	**Arginine**
H	**(His)**	**Histidine**

negatively charged

Q	**(Gln)**	**Glutamine**
D	**(Asp)**	**Aspartic acid**

10 Publications

Publications in Journals

Opatz S, Polzer H, Herold T, **Konstandin NP**, Ksienzyk B, Zellmeier E, Vosberg S, Graf A, Krebs S, Blum H, Hopfner KP, Kakadia PM, Schneider S, Dufour A, Braess J, Sauerland MC, Berdel WE, Büchner T, Woermann BJ, Hiddemann W, Spiekermann K, Bohlander SK, Greif PA. **Exome sequencing identifies recurring FLT3 N676K mutations in core-binding factor leukemia.** *Blood. 2013 Sep 5;122(10):1761-9. doi: 10.1182/blood-2013-01-476473. Epub 2013 Jul 22.*

Neumann M, Heesch S, Schlee C, Schwartz S, Gökbuget N, Hoelzer D, **Konstandin NP**, Ksienzyk B, Vosberg S, Graf A, Krebs S, Blum H, Raff T, Brüggemann M, Hofmann WK, Hecht J, Bohlander SK, Greif PA, Baldus CD. **Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations.** *Blood. 2013 Jun 6;121(23):4749-52. doi: 10.1182/blood-2012-11-465138. Epub 2013 Apr 19.*

Greif PA, **Konstandin NP**, Metzeler KH, Herold T, Pasalic Z, Ksienzyk B, Dufour A, Schneider F, Schneider S, Kakadia PM, Braess J, Sauerland MC, Berdel WE, Büchner T, Woermann BJ, Hiddemann W, Spiekermann K, Bohlander SK. **RUNX1 mutations in cytogenetically normal acute myeloid leukemia are associated with a poor prognosis and up-regulation of lymphoid genes.** *Haematologica. 2012 Dec;97(12):1909-15. doi: 10.3324/haematol.2012.064667. Epub 2012 Jun 11.*

Greif PA, Dufour A, **Konstandin NP**, Ksienzyk B, Zellmeier E, Tizazu B, Sturm J, Benthaus T, Herold T, Yaghmaie M, Dörge P, Hopfner KP, Hauser A, Graf A, Krebs S, Blum H, Kakadia PM, Schneider S, Hoster E, Schneider F, Stanulla M, Braess J, Sauerland MC, Berdel WE, Büchner T, Woermann BJ, Hiddemann W, Spiekermann K, Bohlander SK. **GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia.** *Blood. 2012 Jul 12;120(2):395-403. doi: 10.1182/blood-2012-01-403220. Epub 2012 May 30.*

Greif PA, Yaghmaie M, **Konstandin NP**, Ksienzyk B, Alimoghaddam K, Ghavamzadeh A, Hauser A, Graf A, Krebs S, Blum H, Bohlander SK. **Somatic mutations in acute promyelocytic leukemia (APL) identified by exome sequencing.** *Leukemia. 2011 Sep;25(9):1519-22. doi: 10.1038/leu.2011.114. Epub 2011 May 24.*

Konstandin N, Bultmann S, Szwagierczak A, Dufour A, Ksienzyk B, Schneider F, Herold T, Mulaw M, Kakadia PM, Schneider S, Spiekermann K, Leonhardt H, Bohlander SK. **Genomic 5-hydroxymethylcytosine levels correlate with TET2 mutations and a distinct global gene expression pattern in secondary acute myeloid leukemia.** *Leukemia. 2011 Oct;25(10):1649-52. doi: 10.1038/leu.2011.134. Epub 2011 May 31.*

Greif PA, Eck SH, **Konstandin NP**, Benet-Pagès A, Ksienzyk B, Dufour A, Vetter AT, Popp HD, Lorenz-Depiereux B, Meitinger T, Bohlander SK, Strom TM. **Identification of recurring tumor-**

specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia. 2011 May;25(5):821-7. doi: 10.1038/leu.2011.19. Epub 2011 Feb 22.*

Talks and Posters

6/2010 Poster	**Whole Transcriptome and Exome high throughput sequencing in CLL patients to discover new CLL related Mutations.** 12. Wissenschaftliches Symposium der medizinischen Klinik und Poliklinik III, Herrsching am Ammersee, Germany
6/2011 Poster	**Cytogenetically normal AML with RUNX1 mutations is characterized by high expression levels of the terminal deoxynucleotidyltransferase gene (*DNTT*).** 16th Congress of the Eurpean Hematology Association (EHA), London, United Kingdom
7/2011 Talk	**The Role of TET Proteins in Myeloid Neoplasia** 13. Wissenschaftliches Symposium der medizinischen Klinik und Poliklinik III, Herrsching am Ammersee, Germany
10/2011 Talk	**Genomic 5-hydroxymethylcytosine levels correlate with *TET2 mutations and* a distinct global gene expression pattern in secondary acute myeloid leukemia.** Jahrestagung der Deutschen, Österreichischen und Schweizerischen Gesellschaft für Hämatologie und Onkologie, Bern, Schweiz
7/2012 Talk	**Identification of tumor-specific mutations in CLL by whole exome sequencing** 14. Wissenschaftliches Symposium der medizinischen Klinik und Poliklinik III, Herrsching am Ammersee, Germany
12/2012 Poster	**Whole exome sequencing of CLL before second-line treatment with Fludarabine and Cyclophosphamide with or without Rituximab (REACH-trial)** 56th ASH Annual Meeting and Exposition

11 Acknowledgments

An dieser Stelle möchte ich mich bei allen Personen bedanken die durch ihre fachliche und/oder persönliche Unterstützung zum Gelingen dieser Arbeit beigetragen haben.

- Ein besonderer Dank geht an meinen Betreuer Herrn Prof. Dr. Stefan Bohlander der mir das Thema dieser Dissertation zur überlassen hat und die Möglichkeit diese in seiner Arbeitsgruppe durchzuführen. Ich danke ihm außerdem für sein engagierte Betreuung und Unterstützung bei wissenschaftlichen Fragestellungen.

- Ich möchte mich außerdem bei Herrn Prof. Dr. Heinrich Leonhardt für die Vertretung meiner Doktorarbeit an der Fakultät für Biologie an der LMU bedanken.

- Bei Herrn Prof. Dirk Eick möchte ich mich für die Übernahme des Zweitgutachtens bedanken.

- Die Arbeit wurde durch das REACH Biomarker Projekt (Hoffmann-La Roche) ermöglicht. Von Seiten der Firma Hoffmann-La Roche stand mir Frau Dr. Kerstin Trunzer immer als kompetente Ansprechpartnerin zur Verfügung. Für die Auswertung der klinischen Daten die mir Hoffmann-La Roche für diese Arbeit zur Verfügung gestellt hat möchte ich mich bei Klaas Veenstra und Claude Berge bedanken.
  Außerdem möchte ich mich bei den Kooperationspartnern am Genzentrum bedanken: Herrn Prof. Dr. Helmut Blum und Dr. Stefan Krebs für die Bereitstellung des Genome Analyzers und des Computer Clusters. Alexander Graf und Andreas Hauser danke ich für die Unterstützung bei meinen ersten Gehversuchen in der Bioinformatik.

- Ich möchte mich bei allen aktuellen und ehemaligen Arbeitskollegen der KKG Leukämie bedanken besonders bei den Kollegen der AG Bohlander für das angenehme Arbeitsklima und die gegenseitige Unterstützung: Anna Vetter, Sayantanee Dutta, Sabrina Opatz, Belay Titazou, Medhanie Mulaw, Dity Sen, Naresh Koneru, Purvi Kakadiya und Marjan Yaghmaie. Ein ganz besonderer Dank geht dabei an Bianka Ksienzyk, für die Durchführung unzähliger Sequenzierreaktionen. Für viele fruchtbare Diskussionen und ein unermüdliches Networking, von dem auch dieses Projekt profitierte, möchte ich mich bei Dr. Philipp Greif bedanken. Bei Sebastian Vosberg bedanke ich mich für bioinformatischen Support.
  Mit Freude denke ich an zahlreichen Aktivitäten in der KKG und auch an die wöchentliche Yoga-Stunde vor dem Laboralltag, für die sich das frühe Aufstehen jedes Mal gelohnt hat.

- Ein großes Dankeschön geht auch an meine Kollegen im Labor für Leukämiediagnostik, im Besonderen an die Mitarbeiterinnen der Molekulargenetik: allen voran Dr. Annika Dufour die mir den Einstieg in das REACH Biomarker-Projekt/Studie erleichtert hat und für die Hilfsbereitschaft in wissenschaftlichen Fragen insbesondere in der CLL; bei Dr. Stephanie Schneider möchte ich mich für die freundliche Aufnahme ins Leukämielabor bedanken; bei Gudrun Mellert und Evelyn Zellmeier für ihre technische Expertise.

- Ein herzlicher Dank gilt auch meinen Eltern und meiner Schwester Vera, die mich zu jedem Zeitpunkt unterstützt haben sowie meinem Freund Andreas Rudolph der immer für mich da war, und aus eigener Erfahrung die Höhen und Tiefen während der Entstehung einer Doktorarbeit nachvollziehen konnte.