Institute of Human Genetics
Medical University of Graz
Harrachgasse 21/8 A-8010 Graz
Head: Univ.-Prof. Dr.med.univ. Michael Speicher

Master's Thesis

Identification of plasma biomarkers of bronchial carcinoma

to achieve the degree of
Master of Science

Author:
Katharina A. Woisetschläger, BSc

Evaluator:
Univ.-Doz Dipl.-Chem. Dr.rer.nat. Marcel Scheideler

Institute of Molecular Biotechnology
University of Technology Graz
Petersgasse 14/V A-8010 Graz

Supervisor:
Univ.-Prof. Dr.med.univ. Michael Speicher

Co-Supervisor:
Univ.-Ass.^{in} Mag.^{a} Dr.^{in} rer.nat Ellen Heitzer

Institute of Human Genetics
Medical University of Graz
Harrachgasse 21/8 A-8010 Graz

May 01, 2014

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

……………………………   ………………………………………………..
date   signature

Abstract

Cancer is a heterogeneous and complex disease. Genetic variations as for example single nucleotide polymorphisms, rearrangements of chromosomes or gains and losses of parts of the chromosome or the whole chromosome affect both, the treatment response and the disease susceptibility. One major goal of cancer medicine is to move from fixed treatment regimes to therapies tailored to a patient's individual tumor. This kind of therapy is called 'personalized medicine' and the identification of biomarkers is a key event for the realization of personalized medicine in cancer. As demonstrated in many studies cell free DNA derived from plasma can be used as 'liquid biopsy' to identify copy number variations (CNVs) in a minimal invasive and sensitive way.

Heitzer *et al.* (2013c) established a method called 'plasma-Seq' to generate whole genome sequencing of plasma DNA by using Illumina's MiSeq instrument. With this method tumor genomes can be analyzed noninvasively at low cost within 3 days. Therefore shotgun libraries were prepared using the TruSeq DNA LT Sample preparation Kit from Illumina. Useable CNV profiles for patients having breast-, colon- and prostate cancer had been received by using this plasma-Seq method.

During this study the plasma cell free DNA (cfDNA) of lung cancer patients was investigated based on the plasma-Seq method, with the aim to identify specific biomarkers. To reconstruct tumor genomes five blood samples from patients having lung cancer and a sum of six follow up samples were analyzed. In addition also a sum of 20 blood samples from 11 woman and 9 men without malignant disease were analyzed as control samples. For all samples the concentration was determined followed by quantitative and qualitative analysis as well as whole-genome sequencing.

As determined in other studies before, the blood of cancer patients possessed also a higher concentration of cfDNA than the blood of the control samples in this study. In addition the detected lengths of the DNA fragments were approximately 170 bp and hence, correspond according to Heitzer *et al.* (2013b) to the length of DNA released from apoptotic cells. However, only for one patient (sample *Lunge2*) gains and losses could be detected in the copy number profile. All other genome profiles showed a balanced characteristic.

On the basis of the results obtained for sample *Lunge2* it can be said that the method plasma-Seq used in this study principally provides the opportunity to analyze tumor genomes from peripheral blood obtained from lung cancer patients in detail. But to evaluate the practicability for clinical implementation further analyses with a larger number of samples have to be performed.

Zusammenfassung

Krebs ist eine heterogene und komplexe Krankheit. Sowohl die Reaktion auf die Behandlung als auch die Anfälligkeit für solch eine Krankheit werden durch genetische Veränderungen wie zum Beispiel Einzelnukleotid- Polymorphismen, Strukturveränderungen von Chromosomen oder durch Zugewinne oder Verluste von Chromosomenteilen oder ganzen Chromosomen beeinflusst. Ein Hauptziel der Krebsmedizin ist es, fixe Behandlungsschemata durch personalisierte Medizin - und damit durch Therapien die auf den individuellen Tumor des Patienten abgestimmt sind - zu ersetzen. Für die Realisierung der personalisierten Medizin spielt die Identifizierung von Biomarkern eine wichtige Rolle. Viele bisher publizierte Studien haben bewiesen, dass zellfreie DNA, die aus dem Plasma gewonnen wird, als 'flüssige Biopsie' für die Identifizierung von Kopienzahlveränderungen verwendet werden kann.

Heitzer *et al.* (2013c) entwickelten die Methode 'Plasma-Seq', um das gesamte Genom nur anhand von Plasma-DNA zu sequenzieren. Diese Methode ermöglicht es, das gesamte Tumorgenom nichtinvasiv, kostengünstig und innerhalb von 3 Tagen zu analysieren, indem mit Hilfe eines Kits von Illumina Shotgun- Bibliotheken erstellt wurden. Bisher wurden mit dieser Plasma-Seq Methode zufriedenstellende Ergebnisse für Patienten mit Brust-, Dickdarm- und Prostatakrebs erzielt.

In dieser Studie wurde die zellfreie Plasma-DNA von Patienten mit Lungenkrebs, basierend auf der bereits publizierten Plasma-Seq Methode, untersucht, um spezifische Biomarker zu finden. Dafür wurde Blut von 5 verschiedenen Patienten und eine Summe von 6 Folgeblutproben untersucht. Zusätzlich wurde das Blut von 20 Personen ohne maligne Erkrankungen als Kontrolle analysiert. Für jede Probe wurde die Konzentration ermittelt, quantitative und qualitative Analysen durchgeführt und das gesamte Genom sequenziert.

Wie auch schon in vorherigen Studien beschrieben, war in den Blutproben der Krebspatienten die Konzentration von zellfreier DNA höher als in den Blutproben der Kontrollpersonen. Die untersuchten DNA Fragmente hatten eine durchschnittlichen Länge von 170 bp und entsprachen somit der Länge von Fragmenten, die laut Heitzer *et al.* (2013b) von apoptotischen Zellen freigesetzt werden. Jedoch konnten nur für einen Patienten (Probe *Lunge2*) Zugewinne bzw. Verluste im Kopienzahlenprofil detektiert werden, alle anderen Genom-Profile waren balanciert.

Basierend auf diesen Ergebnissen kann gesagt werden, dass die Methode Plasma-Seq prinzipiell geeignet ist, das aus dem Blut von Lungenkrebspatienten gewonnene Tumorgenom im Detail zu untersuchen, jedoch noch weitere Untersuchungen mit einer höheren Probenzahl notwendig sind, um die Anwendbarkeit im klinischen Bereich abschätzen zu können.

Acknowledgment

At first I want to thank my advisors Univ.-Ass. Mag. Dr.rer.nat Ellen Heitzer and Univ.-Doz Dipl.-Chem. Dr.rer.nat. Marcel Scheideler for their support during the whole work who helped me with words and deeds in many situations. I also want to thank Univ.-Prof. Dr.med.univ. Michael Speicher and Assoz. Prof. Priv.-Doz. Dr.med.univ. Jochen Bernd Geigl who made my work at the Institute of Human Genetics possible. A special thank goes to Univ.-Prof. Dr.med.univ. Horst Olschewski and his team of Pulmonology, LKH Graz for organizing the blood samples needed for my work. Last but not least a big thank goes to my family, my fiancè and friends who supported me during the whole study.

Table of content

1 Introduction

1.1 Human genome

A genome usually contains the biological information and exists in every organism. The human genome is as most genomes made of DNA whereas for example viruses have a genome made of RNA. This human genome is composed of two distinct parts, the nuclear genome and the mitochondrial genome which is a circular DNA molecule of 16.569 nucleotides. In contrast, the haploid nuclear genome consists of 3.200.000.000 nucleotides divided in 23 linear molecules and each of these molecules is contained in a different chromosome. The majority of the cells in human body (in sum 10^{13} cells), the somatic cells, are diploid and contain two copies of each autosome (chromosome 1-22) and two sex chromosomes (XY for male and XX for female). Only the gametes or sex cells are haploid and have only 23 chromosomes, only one copy of each [Brown (2002); Strachan, Read (1999)].

1.1.1 DNA

DNA is an unbranched, linear polymeric molecule which is composed of smaller molecules called nucleotides, which can be linked together to form chains of hundreds, thousands or millions units in length [Brown (2002), Nature education (2014)]. Nearly all living cells contain DNA but each organism has a unique sequence of DNA because the arrangement of the nucleotides differs among individuals [Nature education (2014)]. In general desoxyribose, a carbon-based sugar molecule, a nitrogenous base (cytosine, thymine (single-ring pyrimidines), adenine or guanine (double-ring purines)) and a phosphate group (one, two or three linked phosphate units) are the three main components of nucleotides. Consequently there exist four different DNA nucleotides (dCTP, dTTP, dATP or dGTP) because of the four different nitrogenous bases. If a molecule consists only of sugar and base it is called nucleoside [Brown (2002); Nature education (2014); Strachan, Read (1999)].

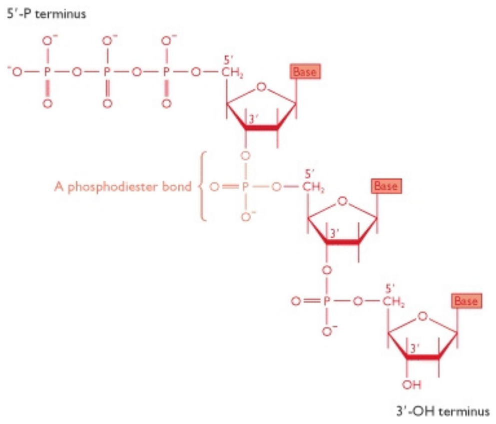


Figure 1: Structure of a polynucleotide [Brown (2002)]

As shown in figure 1 individual nucleotides are linked together by a phosphodiester bond to form a polynucleotide. There exist two ends of the polynucleotide: the 3’ OH terminus where an unreacted hydroxyl is attached to the 3’ carbon and the 5’ P terminus where an unreacted triphosphate group is attached to the 5’ carbon. Consequently two chemical directions result: a 5’→3’ or 3’→5’. For example a 5’→3’ synthesis is done by all natural DNA polymerase [Brown (2002); Nature education (2014)].

Via hydrogen bonds, the nitrogenous bases of one polynucleotide are linked to the nitrogenous bases of the other polynucleotide to form a right-handed double helix like shown in figure 2 [Nature education (2014)]. Adenine can only base-pair with thymine and cytosine can only base-pair with guanine [Brown (2002)]. The polynucleotides used for the double helix are ‘upside down’ which mean that the two strands run in opposite directions [Brown (2002); Nature education (2014)].

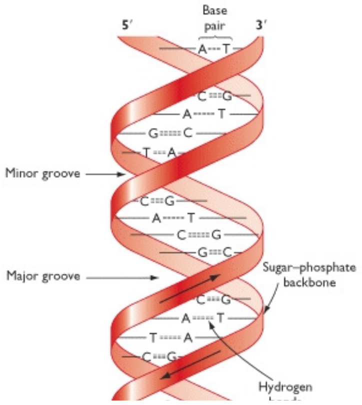


Figure 2: DNA double helix [Brown (2002)]

The DNA of one single cell has a length of approximately two meters when arranging the DNA in a single straight piece. Therefore, the DNA has to be packed to fit in a cell like shown in figure 3. Eukaryotic DNA for example is wrapped around special proteins called histones and the DNA and the histone proteins together are called chromatin. Via supercoiling, a twisting process, the DNA can be further compressed. At the end this compressed DNA is arranged into chromosomes [Nature education (2014)].

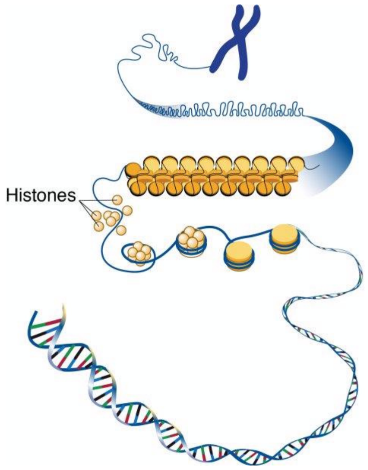


Figure 3: DNA packing [figure modified from National Human Genome Research Institute (2010)]

A gene is the part of the human genome that contains biological information. One or more protein molecules are specified by one gene [Brown (2002); Vogelstein *et al.* (2013)]. In general the expression of a gene is a one-way system, described as the central dogma of molecular biology, in which the DNA is transcript into RNA and this RNA is translated into a protein [Strachan, Read (1999)]. Therefore messenger RNA (mRNA) is used which directs the synthesis of the protein coded by the gene. The amount of protein-coding genes is only approximately 1.5% of the total genome (approximately 25.000 genes) [Brown (2002); Vogelstein *et al.* (2013)]. The rest is called ‘dark matter’ of the genome and according to Gibb et al.(2011) ‘may play a major biological role in cellular development and metabolism’ [Gibb et al. (2011)].

1.1.2 DNA mutations

The human genome is not a static unit but changes over the time due to larger scale rearrangements caused by recombination and small-scale sequence alterations caused by mutation [Brown (2002); Strachan, Read (1999)]. Recombination and mutation are unrelated and have to be distinguished although both effects result in changes to the genome [Brown (2002)].

A mutation is defined as a 'change in the nucleotide sequence of a short region of a genome' [Brown (2002)]. There exist different kinds of mutation: First, the **point mutation** is a very common mutation where only one nucleotide is replaced by another. The other two mutations involve **insertion** or **deletion** of one or more nucleotides [Brown (2002); Strachan, Read (1999)]. On the one hand mutations can be caused by errors in DNA replication, on the other hand they can be caused by mutagens like the natural ionizing radiation or chemicals which change the structures of individual nucleotides [Brown (2002); Strachan, Read (1999)]. To minimize the number of mutations two kinds of DNA-repair enzymes exist: The pre-replicative enzymes check the DNA and replace the nucleotides with unusual structures before the replication takes place. The post- replicative enzymes in contrast search for errors in the newly synthesized DNA [Brown (2002)]. Nevertheless the frequency for an uncorrected replication error per incorporated nucleotide is about 10^{-9}-10^{-11} [Strachan, Read (1999)].

Recombination is the restructuring of parts of a genome, for example the transposition of a genome part from one position to another within a chromosome or between two chromosomes, or the exchange of segments of homologous chromosomes during meiosis [Brown (2002)].

The presence of mutation as well as recombination can have dramatic effects on the cell. In the case of a mutation in a key gene, the protein coded by the mutated gene can be defective and causes the cell to die. However, many mutations or recombinations have a less impact on the phenotype of a cell and some have none at all. It is to mention, that events that are not lethal can contribute to the evolution of the genome in case they are inherited during the reproduction of the organism [Brown (2002); Strachan, Read (1999)].

1.1.3 Free circulating tumor DNA

In clinical oncology genetic and molecular biomarkers play an important role. Via biomarkers a detection of the disease at an early stage, prediction who will develop cancer and the selection of targeted therapies is possible [Cima *et al.* (2011); Ziegler *et al.* (2012)]. For biomarkers, different definitions can be found in literature. The most convenient definition was developed by the Biomarkers Definitions Working Group (2001): 'A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacologic responses to a therapeutic intervention'. Three main types of biomarkers: DNA biomarker, DNA tumor biomarker and general biomarker. Examples for DNA biomarkers are single nucleotide polymorphisms (SNP), deletions, insertions, short tandem repeats (STRs) or other variation on the DNA sequence level. *EGFR* or *K-RAS* are examples for predictive DNA tumor biomarkers and play an important role in patients with advanced non-small-cell lung cancer. All other forms of biomarkers such as protein, RNA or metabolite measurement, measured in for example tissues or cell lines are called 'general biomarkers' [Ziegler *et al.* (2012)]. In many types of cancer cell free DNA (cfDNA) is present in high concentration in serum or plasma and can be used as a genetic biomarker [Ghorbian, Ardekani (2012)].

The existence of cell free nucleic acid (cfNA) in human blood was described for the first time in 1948 by Mandel and Métais [Ghorbian, Ardekani (2012)]. Many current studies have shown that the blood of cancer patients possess a high concentration of cfNA. During tumor progression a release of both, normal (wild-type) DNA and tumor-derived DNA into the blood stream can occur [Schwarzenbach *et al.* (2011)]. These cfDNA in plasma or serum can for example be used as a noninvasive 'liquid biopsy' to identify prognostic, predictive and pharmacokinetic biomarkers [Dawson *et al.* (2013); Heitzer *et al.* (2013a); Ziegler *et al.* (2012)]. For personalized medicine a differentiation between predictive and prognostic biomarkers is important: predictive biomarkers are connected with the response to a treatment, whereas prognostic biomarkers are used for predicting the progress of the disease [Ziegler *et al.* (2012)]. The use of cfDNA as a liquid biopsy avoids the need for tumor tissue biopsies and make the collection of repeated blood samples possible [Schwarzenbach *et al.* (2011)]. A challenge of getting tumor tissues can be the elaboration of procedures to acquire high-quality metastatic tissue for analysis. In addition this material cannot be used for biomarker detection for personalized treatment decisions but only for research purposes and only provides a snapshot of mutations at a given time and location [Forshew *et al.* (2012); Heitzer *et al.* (2013c)]. This makes 'liquid biopsies' very important in future; due to the possibility to collect repeated samples the changes in cfDNA during cancer treatment and during the natural development of the disease can be investigated and assessment of cancer patients after therapy is possible [Schwarzenbach *et al.* (2011); Ghorbian, Ardekani (2012)]. The extraction and analyses of cfDNA have a major clinical utility as tumor-related genetic and epigenetic alterations can be determined [Schwarzenbach *et al.* (2011)]. Furthermore, it is possible to reconstruct complex tumor genomes from the peripheral blood of patients with cancer [Heitzer *et al.* (2013b)].

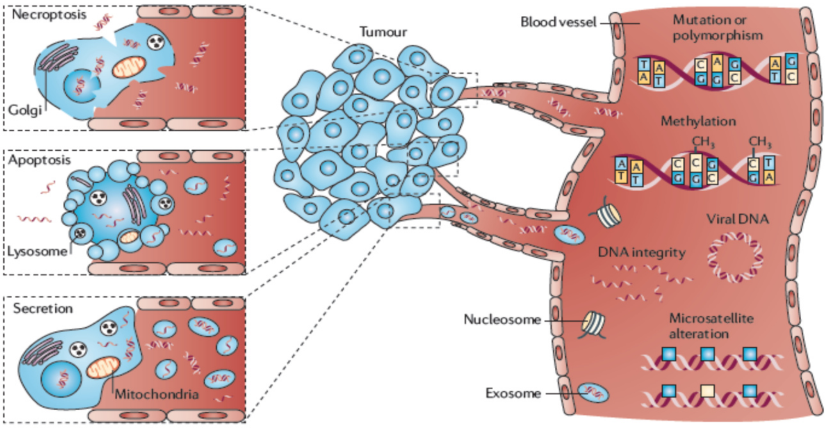


Figure 4: cfDNA release into blood [Schwarzenbach *et al.* (2011)]

In literature there are different suggestions how cfNA can be released into blood. Schwarzenbach et al. (2011) mentioned previously apoptosis, necrosis and secretion as possible causes for such release (figure 4). Usually macrophages or other scavenger cells phagocytize apoptotic and necrotic cells [Schwarzenbach *et al.* (2011)]. Also Ghorbian and Ardekani (2012) mentioned apoptosis and necrosis of the cells as possible reasons for cfNA release. Just as well as the detachment of cells into the blood stream [Ghorbian, Ardekani (2012)]. Digested DNA can be released either single-stranded or double-stranded, with a size between about 100 bp and 21 kilobases, into the tissue environment by magrophages that engulf necrotic cells [Schwarzenbach *et al.* (2011)]. According to Heitzer *et al.* (2013b) the length of DNA released from apoptotic cells is within a size range of 85-230 bp whereas DNA fragments from necrotic cells have sizes larger than 10.000 bp [Heitzer et al. (2013b)]. But till now the exactly physiology and rate of release is still not well understood [Schwarzenbach *et al.* (2011); Heitzer *et al.* (2013c)]. Factors like tumor cell proliferation and tumor burden may influence these events. It is to mention that cfNA originates not only from the primary tumor, also blood circulating tumor cells and micrometastasic deposits can contribute to the release of cfNA. Scientists estimated that up to 3.3% of tumor DNA may enter the blood every day if a patient has a tumor that weights 100g (about 3 x 10^3 tumor cells). Thus, the size and also the state of the tumor influence the proportion of cfDNA that originates from tumor cells. Other influences on the amount of cfDNA are degradation, clearance and other physiological filtering events of the blood. Because of these events, cfDNA is cleared from the circulation having a variable half-life in the circulation between 15 minutes and several hours. Hence, the concentration of overall cfDNA varies obviously in plasma of healthy control donors as well as cancer patients [Schwarzenbach *et al.* (2011)]. In addition the plasma cfDNA level can be increased due to trauma, premalignant states, and inflammation, in patients suffering from illnesses and after exercise [Ghorbian, Ardekani (2012)]. These are reasons why the amount of circulating DNA as a diagnostic value has been called into question by many studies. Heitzer *et al.* (2013b) determined in their work that not all patients with metastatic disease release a measurable quantity of tumor DNA into the circulation and mentioned as an explanation that maybe because of the short half-life of plasma DNA no significant amount of tumor DNA had been released into the circulation before blood drawing [Heitzer et al. (2013b)].

1.1.4 DNA analysis

The toolkit of techniques used for studying DNA molecules was generated during the 1970s and 1980s. In the early 1970s the development of these techniques was stimulated by the application of enzymes which manipulate DNA molecules in test tubes. These enzymes were used to copy DNA molecules and to cut DNA molecules into shorter fragments which are then used to create new combinations. These manipulations provided a basis for gene cloning [Brown (2002)]. In the mid-1980s the polymerase chain reaction (PCR), a further technical breakthrough, was invented by Mullis [Brown (2002); Strachan, Read (1999)]. The procedure of this PCR and the derived qPCR are presented in the following chapters.

1.1.4.1 PCR

The polymerase chain reaction is used for specific in vitro production of multiple copies of defined DNA sequences and is implemented in many areas like for example the forensic laboratory, virology and cancer therapy [Lynch, Brown (1990)]. An advantage of this technique is the extremely high sensitivity and specificity. Hence PCR is able to work with minuscule amounts of starting DNA [Brown (2002); Lynch, Brown (1990); Strachan, Read (1999)]. So, the sequences of DNA present in hair or bloodstains can be obtained using PCR, to give two examples. In addition it is easy to obtain products via PCR that represent a single segment of the genome from a number of different DNA samples. Hence, PCR can be used to screen human DNA samples for mutations associated with genetic diseases [Brown (2002)]. However, a major limitation of PCR is the need to know the boundary regions of the DNA to be amplified [Lynch, Brown (1990)]. This means it is not possible to purify parts of a genome by PCR without having prior DNA sequence information [Brown (2002); Strachan, Read (1999)]. Furthermore a DNA with the length of only 5 kb can be copied without much difficulty. With modification of the standard technique amplification up to 40 kb is possible but fragments with a length greater than 100 kb - which are important for the genome sequencing projects - are unattainable by PCR [Brown (2002)].

To perform a PCR reaction target DNA, a thermostable DNA polymerase (for example Taq polymerase), a supply of nucleotides and a pair of oligonucleotide primers are mixed together [Brown (2002)]. The primers, which often have a length of about 15-25 nucleotides, have to bind to the target DNA at both sides surrounding the segment that needs to be copied [Brown (2002); Strachan, Read (1999)]. Therefore the sequences of the attachment sites have to be known before [Brown (2002)]. Repeated cycles of heat denaturation, annealing and primer extension are involved in

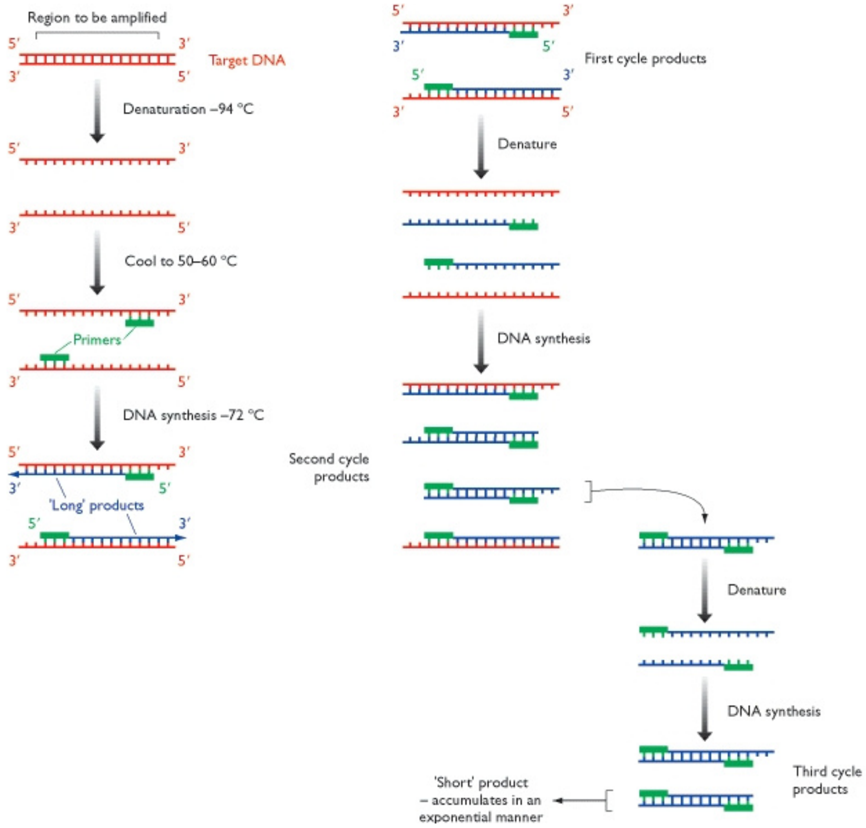


Figure 5: Process of polymerase chain reaction [Brown (2002)]

the PCR procedure and after each cycle, the amount of DNA is ideally doubled [Lynch, Brown (1990)]. In general, in the first step of the reaction the mixture is heated to 93-95°C to denature the target DNA into single-stranded molecules (figure 5). Then the temperature is reduced to 50-70°C (depending on the melting temperature of the expected duplex) to attach the primers, that are present in vast molar excess, to their complementary annealing position [Brown (2002); Lynch, Brown (1990); Strachan, Read (1999)]. At this temperature also rejoining of the single strands of the target DNA can occur. To start the DNA synthesis, the temperature is raised to 70-75°C, the perfect condition for Taq polymerase to work [Brown (2002); Strachan, Read (1999)]. The primers are extended by incorporating nucleotides by the DNA polymerase [Lynch, Brown (1990)]. As shown in figure 5 a long product is synthesized from each of the two DNA strands. In the next cycle, these long products will serve as new templates for another DNA synthesis [Brown (2002)]. The primers themselves become incorporated into the fragments and hence the termini of the short fragments are defined by these primers [Lynch, Brown (1990)]. After 30 denaturation-annealing-synthesis-cycles over 250 million blunt ended short products derived from each starting molecule. There exist different ways to determine the results of PCR but usually the products are analyzed by agarose gel electrophoresis or sequencing [Brown (2002); Strachan, Read (1999)].

1.1.4.2 Real time PCR

Real time PCR is a modification of the standard PCR method and is used for quantification of DNA. In real time PCR the synthesis of the products is measured over time as it occurs [Brown (2010)]. The amount of amplified product present at any cycle is proportional to the amount of starting DNA. Therefore, the product yield will be greater the more DNA is present in the starting mixture. For quantification of the test DNA the fluorescence intensity is compared to the intensity of a control DNA which starting amount is known. For the comparison the cycle at which the amount of PCR product reaches a pre-set threshold is identified. The bigger the amount of starting test DNA is, the earlier this threshold is reached (figure 6) [Brown (2010)].

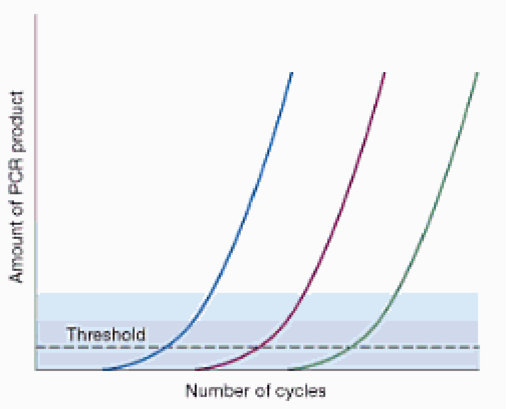


Figure 6: Quantification by real time PCR [Brown (2010)]

There exist two strategies to follow the product synthesis in real time: i) using a dye causing a fluorescent signal when binding to a double-stranded DNA or ii) a short oligunulceotide (reporter probe) which gives a fluorescent signal when binding to a PCR product [Brown (2010)]. An example for an often used dye is SYBR- Green. SYBR- Green is a dye preferentially binding to the minor groove of double-stranded DNA [QIAGEN SABioscience (2008); Dorak (2006)]. The intensity of the resulted fluorescent emission signal is proportional to the amount of present double-stranded DNA. Although, non-specific increases in the signal can occur if primers anneal to one other, SYBR Green is often used because of its high detection sensitivity and its relatively low cost [QIAGEN SABioscience (2008)]. In contrast, by using reporter probe, the inaccuracies caused by primer-primer annealing are less because the probe only binds to the PCR product. The design of the used oligonucleotide is following: One end of the oligonucleotide including a fluorescent dye base pair to the other end including a quenching compound, which inhibits the fluorescent signal. By hybridization between the oligonucleotide and the PCR product the quencher moves away from the dye and fluorescent signal can be generated [Brown (2010)].

In figure 7 the incorporation of SYBR Green is represented graphically. In an analogues manner compared to PCR, the DNA is denatured in the first step. The SYBR Green dye does not bind to single stranded or denatured DNA and a free dye has a very low fluorescence. A double-stranded DNA is formed during primer annealing and primer extension by Taq DNA Polymerase and SYBR Green gets incorporated in the double -stranded DNA causing a dramatic increase in fluorescent signal. The fluorescent signal increases proportionally to the number of incorporated SYBRE Green molecules. In each cycle this process is repeated causing an increase of total fluorescence that is measured constantly [Dorak (2006)].

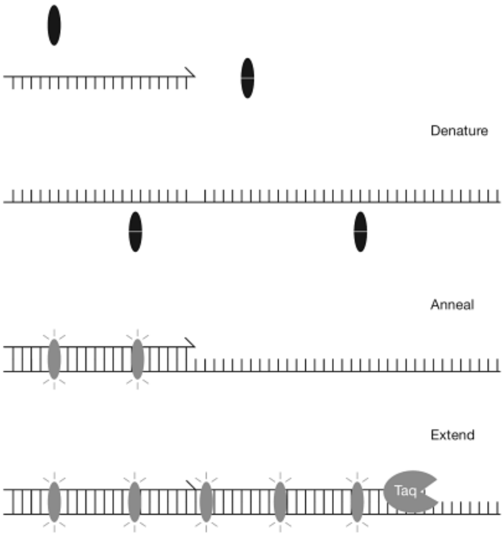


Figure 7: Real time PCR using SYBR Green [Dorak (2006)]

1.2 Cancerogenesis

Cancer is a disease which is characterized by uncontrolled growth and spread of abnormal cells [Cancer Facts and Figures (2012)]. Usually, normal cells die in an orderly fashion but cancer cells continue to grow and form abnormal cells instead of dying [American Cancer Society (2014)]. The reason for the uncontrolled growth and the spread of abnormal cells is the dysfunction of genes that control cell growth and division. Both internal factors, like for example mutations and hormones, and external factors, like for example tobacco and chemicals, are reasons for the genetic damage [Cancer Facts and Figures (2012)].

1.2.1 Genetic alterations and characteristics of tumors

Cancer involves changes to the DNA at the cellular level and via sequencing efforts over the past decade the genomic landscapes of common forms of human cancer could be obtained [Ziegler *et al.*

(2012); Vogelstein *et al.* (2013)]. A human cancer comprises of a small number of genes altered in a high percentage and much more genes that are altered only infrequently. According to studies, approximately 140 genes can 'drive' tumorigenesis, when altered by intragenic mutations. In general, two to eight of these 'driver gene' mutations, developed over the course of 20 to 30 years, cause a common tumor and the remaining mutations are so- called 'passenger mutations' [Vogelstein *et al.* (2013)]. According to Bozic *et al.* (2010) a 'driver gene mutation' is a mutation that confers a selective growth advantage to the cell in which it occurs. A 'passenger mutation', however, is a mutation that has no effect on the selective growth advantage of the cell but occur in a cell that subsequently or coincidentally acquire a driver mutation [Bozic *et al.* (2010)]. It is to mention, that there is a difference between 'driver gene' and a 'driver gene mutation'. A driver gene contains driver gene mutations but may also contains passenger gene mutations [Vogelstein *et al.* (2013)].

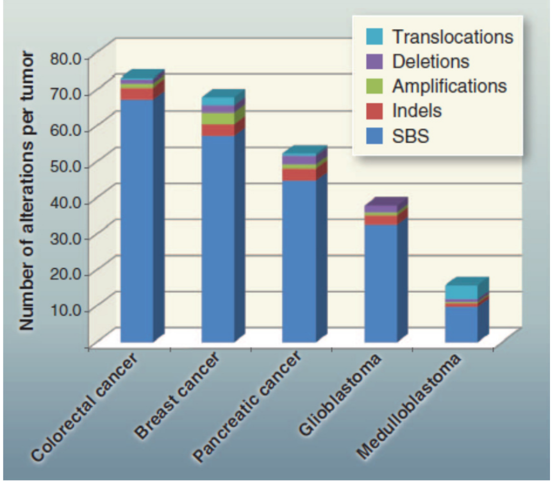


Figure 8: Alterations affecting protein-coding genes in different tumor types [Vogelstein *et al.* (2013)]

The general amount of mutated genes in a tumor differs depending on the type of tumor. Tumor types such as colon-, brain-, breast- and pancreas tumor display an average of 33-66 mutated genes of which approximately 95% are single- base substitutions (SBS) whereas the rest are small insertions and deletions (indels), translocations, homozygous deletions and amplifications (figure 8) [Vogelstein *et al.* (2013)].

In comparison to normal cells, tumor cells have the same rate of point mutations but there is an elevation of the rate of chromosomal changes in cancer [Vogelstein *et al.* (2013)]. Therefore, most solid tumors are characterized by large- scale structural genomic rearrangements and by abnormal chromosome numbers (aneuploidy) [Orr, Compton (2013); Vogelstein *et al.* (2013)]. Changes like chromosomal translocations and aneuploidy usually arise in individual somatic cells [Clancy (2008)]. A further characteristics of cancer is chromosomal instability (CIN) a cause of tumor aneuploidy [McGranahan *et al.* (2012)]. The difference between aneuploidy and chromosomal instability (CIN) is defined as following: 'Aneuploidy is a state of abnormal chromosome number and all the cells in the tumor have the same defective karyotype'. In contrast, 'CIN is a higher rate of chromosome mis-segregation that enhances karyotypic diversity in cells within the same tumor' [Orr, Compton (2013)]. Important to mention is that aneuploidy does not necessarily involve CIN but CIN causes aneuploidy all the time [McGranahan *et al.* (2012)].

Figure 9 shows the difference between the two types of chromosomal instability: i) numerical CIN where gain and loss of whole chromosomes occur and ii) the structural CIN, including the increased rate of structurally abnormal chromosomes (for example gains and losses of chromosome fragments or translocations). The CIN status can be determined by various methods for example fluorescence in situ hybridization (FiSH), comparative genomic hybridization (CGH) or next generation sequencing [McGranahan et al. (2012)]. The latter will be discussed in chapter 1.3.1.

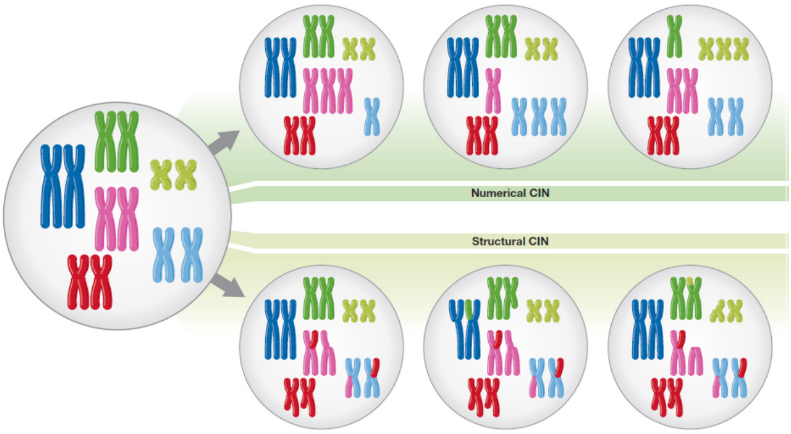


Figure 9: Numerical and structural chromosomal instability [McGranahan *et al.* (2012)]

A further characteristic of a tumor is its heterogeneity. According to Vogelstein *et al.* (2013) there exist four relevant types of heterogeneity relative to tumorigenesis: i) the intratumoral heterogeneity, ‘a heterogeneity among the cells of one tumor’, ii) the intermetastatic heterogeneity, a ‘heterogeneity among different metastatic lesions of the same patient’, iii) the intrametastatic heterogeneity, a ‘heterogeneity among the cells of an individual metastasis’, and iv) the interpatient heterogeneity, a ‘heterogeneity among the tumors of different patients’. These types of heterogeneity and especially the interpatient heterogeneity complicate the design of a uniformly effective treatment and hence the research in the field of ‘personalized medicine’ is of great importance [Vogelstein *et al.* (2013)].

1.2.2 Mutation timing and hallmarks of cancer

Tumors acquire a series of mutations over time and evolve from benign to malignant lesions. The first mutation, so-called ‘gatekeeping’ mutation increases the grow rate of the normal epithelial cell and allow the cell to outgrow the surrounding cells. This first mutation causes a small adenoma with a slow growth but an additional mutation in another gene (for example *KRAS*) leads to a further round of clonal growth with an expansion of cell number. Further mutations in other genes such as *TP53* or *SMAD4* can generate a malignant tumor with the possibility to metastasize to lymph nodes and distant organs. In certain tumors of self-renewing tissues the number of mutations is directly correlated with age and hence, for example a 90-year-old patient has compared to a 45-year old patient with the morphologically identical colorectal tumor twice as many mutations [Vogelstein *et al.* (2013)].

Hallmarks of cancer, which are basically genome instabilities, comprise six capabilities acquired during the multistep development of human tumors and cause the genetic diversity that accelerates their acquisition and inflammation. Hanahan, Weinberg (2011) defined them as following: evading growth suppressors, resisting cell death, sustaining proliferative signaling, inducing angiogenesis, enabling replicative immortality and activating invasion and metastasis. Those six hallmarks play an important role in the transformation of normal cell into malignant and enable tumor growth and metastatic dissemination (figure 10). In the last decade two further hallmarks, reprogramming of energy metabolism and evading immune destruction, were added to the list of hallmarks [Hanahan, Weinberg (2011)].

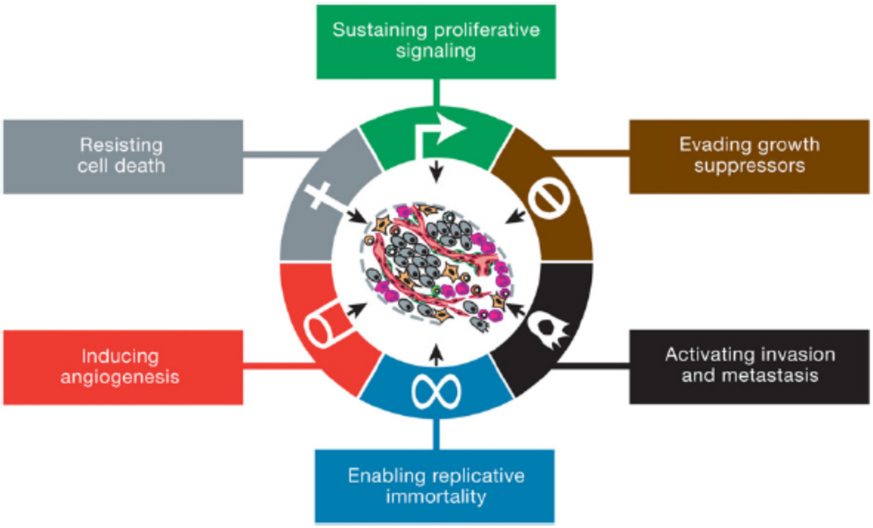

Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality

Figure 10: Hallmarks of cancer [Hanahan, Weinberg (2011)]

Normal cells regulate the cell growth and cell division cycle by the production and release of growth-promoting signals. In contrast, cancer cells have the ability to **sustain proliferative signaling** in different ways: On the one hand they can produce growth factor ligands themselves and regulate the autocrine proliferative stimulation with the expression of cognate receptors. On the other hand cancer cells can stimulate normal cells within the surrounding stoma to get supplied with growth factors [Hanahan, Weinberg (2011)].

In addition powerful programs that negatively regulate cell proliferation have to be circumvented by cancer cells (**evading growth suppressors**). Most of the time so-called tumor suppressor genes play an important role in these programs. RB (retinoblastoma- associated) and TP3 proteins are two examples for proteins encoded by tumor suppressor genes. The RB protein decides whether a cell should continue its growth- and division- cycle or not. The TP3 can, depending on the degree of damage to the genome, stop the cell-cycle progression until the damage is removed or trigger apoptosis. Consequently an inactivation of these tumor suppressor genes plays an important role in tumor development [Hanahan, Weinberg (2011)].

To limit or circumvent apoptosis, tumor cells develop different strategies (**resting cell death**). As mentioned before a common strategy is the loss of TP53 tumor suppressor function. But also downregulating proapoptotic factors, or short-circuiting the extrinsic ligand-induced death pathways and as a result of this the increased expression of antiapoptotic regulators (Bcl-2) or of survival signals are possibilities to limit or circumvent apoptosis [Hanahan, Weinberg (2011)].

Normal cells undergo only a limited number of successive cell growth-and-division cycles but in contrast cancer cells possess unlimited replicative potential (**enabling replicative immortality**). This transition is called 'immortalization', a process to proliferate without crisis (involves cell death) and senescence (a non-proliferative but viable state). The telomeres, composed of tandem hexanulceotide repeats, protect the ends of chromosomes and are shortened with every cell division. Telomere repeat segments are added to the ends of telomeric DNA by the enzyme telomerase, a specialized DNA polymerase. This telomerase is almost absent in normal cells but expressed in spontaneously immortalized cells, for example cancer cells. Consequently, the telomerase counter the progressive telomere erosion and leads to a resistance to the induction of senescence and crisis [Hanahan, Weinberg (2011)].

As well as normal tissue, also tumors need nutrients and oxygen for survival and an ability to remove metabolic waste and carbon dioxide. **Angiogenesis**, the term for spreading of vessels, addresses these needs and is normally part of physiologic processes such as female reproductive cycling and wound healing. However, during tumor progression this angiogenic process is always active, causing continually sprout of new vessels which help to satisfy the demand of proliferation and continuous cell growth [Hanahan, Weinberg (2011)].

These vessels are part of the multistep process of **invasion and metastasis** which consists of discrete steps. The first step of this multistep process is the local invasion which is followed by the intravasation by cancer cells into neighboring blood and lymphatic vessels. Afterwards the cancer cells are transported through these vessels and escape into the parenchyma of distant tissues. Then micrometastases, small nodules of cancer cells, are formed which grow into macroscopic tumors (colonization) [American Cancer Society (2014); Hanahan, Weinberg (2011)]. Millions of cells are released into the circulation from an advanced tumor every day but because of the short half-lives only a miniscule fraction cause a metastatic lesion [Vogelstein *et al.* (2013)]. Furthermore, each kind of tumor has its own metastasis-characteristics: The lung cancer for example prefers the liver, central nervous system, the bones and the adrenal gland as sites for metastasis [Bast Jr, Robert C. *et al.* (2000)].

1.2.3 Proto- oncogenes and tumor- suppressor genes

As mentioned before in tumor initiation and progression the activation of cellular oncogenes and the inactivation of tumor suppressor genes are critical steps. Because of their importance they will be described more detailed in this chapter. Oncogenes regulate abnormal cell proliferation whereas tumor suppressor genes inhibit cell proliferation and tumor development. These genes are lost or inactivated in many tumors [Cooper (2000)].

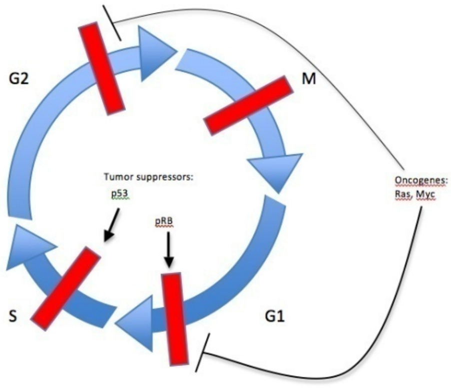


Figure 11: Cell cycle control by tumor suppressors and oncogenes [Chow (2010)]

Figure 11 shows how tumor suppressors and oncogenes control the cell cycle including G (gap) phase, chromosome replication (S phase) and chromosome segregation (mitosis). The tick red bars represent the checkpoints. In normal cells proto-oncogenes control this process but point mutation, translocation or amplification can alter them to oncogenes and consequently their abnormal protein products show increased activity that contributes to tumor growth. A normal cell should stop within a G phase but a tumor cell run through the whole cell cycle which leads to uncontrolled cell division. Additionally there is the possibility that oncogenes rescue cells from apoptosis. The protein Ras for example is a molecular switch that is depending on the form of nucleotide to which it is bound, turned on or off. A mutation which converts the proto-oncogene to an oncogene can render Ras permanently active and thus can leads to serious consequences [Chow (2010)].

Cellular growth and division or even apoptosis is restricted via proteins coded from tumor suppressor genes. In contrast to oncogenes, tumor suppressor genes are the 'brake' of the cell cycle. In figure 11 pRb (known as retinoblastoma protein) which corresponds to the gene *RB1*, the first identified tumor suppressor gene, is shown. Another important tumor suppressor gene in the human body is the *p53* gene which is essential for maintaining the G1 to S cell cycle checkpoint. An inactivation of a tumor suppressor gene like this allows an uncontrolled cell division. It is important to mention that for tumorigenesis to occur, both paternal and maternal 'copies of a gene coding for a tumor suppressor must usually be altered' [Chow (2010)].

Important oncogenes and tumor suppressor genes detected in lung carcinoma are described in chapter 1.4.3.3.

1.3 Copy number variation

The genetic information in a human cell is encoded in 46 chromosomes. Single- nucleotide variations (SNVs) and copy number variations (CNVs) are variations occurring in these chromosomes and are amongst other factors the driving forces in cancer [Zong *et al.* (2012)].

Genetic variations in the human genome range from single nucleotide changes to large, microscopically visible chromosome anomalies. Copy number variations (CNVs) or copy number polymorphisms (CNPs) are for example deletions, insertions, duplications and complex multi-site variations and are found in all humans [Redon *et al.* (2006)]. A CNV was defined by Redon *et al.* (2006) as a 'DNA segment of one kilobase (kb) or larger that is present at a variable copy number in comparison with a reference genome'. Conrad *et al.* (2010) detected that the CNV size ranges from 443 bp to 1.28 Mb with a median size of 2.9 kb and there are 1098 CNVs when comparing two genomes by CGH [Nature education (2014)]. Results from the HapMap Project showed that CNVs cover approximately 360 megabases which equates 12% of the human genome. Not all CNV cause a phenotype but most of them are linked with disease. Redon and colleagues found out 'that CNVs typically lie outside of coding sequences and ultraconserved regulatory elements', sequences of at least 200 bp that are entirely conserved across several species. Genes involved in cell adhesion, responses to chemical stimuli and the sensory perception of smell are these functional categories with the greatest enrichment for CNVs [Clancy (2008)].

1.3.1 Methods for detection

The analysis of genome copy number variations at the single-cell level is important for, amongst others, clinical applications (non-invasive prenatal diagnosis and pre-implantation) and studies of tumor heterogeneity. In general, this analysis can be done by following methods: Fluorescence in situ hybridization (FISH), array- comparative genome hybridization (CGH) and next generation sequencing. However, the resolution and scope of FISH are severely limited and the method is time consuming [Fiegler *et al.* (2007)].

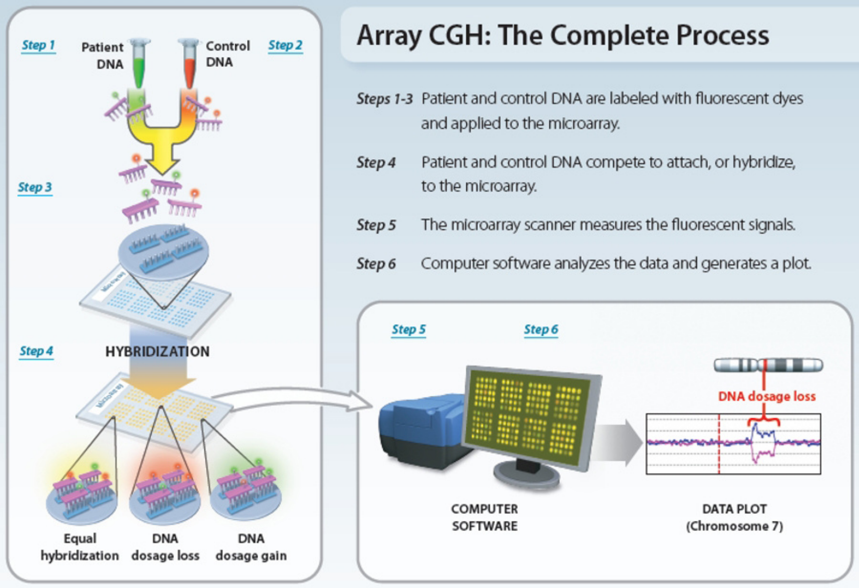
Because of this thesis the attention of this chapter will be on next generation sequencing and the method array CGH will only be described for the sake of completeness.

1.3.1.1 Array CGH

Array CGH combines the principles of CGH (comparative genomic hybridization) with the use of microarrays and is used to monitor genetic variation like gains and losses with an, until then, unprecedented accuracy for somatic tissues [O'Huallachain *et al.* (2012); Theisen (2008)]. Rearrangements as small as about 10 kb can be detected by array CGH. These arrays also can be used for the detection of mosaic events in tissues [O'Huallachain *et al.* (2012)].

Array CGH, which is based on DNA microarrays, detects the genetic variations by comparison with a normal reference DNA and helps to find genes involved in cancer progression because it is assumed that oncogenes are located in gained regions and tumor suppressor genes in lost regions. Furthermore array CGH should improve the classification of tumors [Hupe *et al.* (2004)].

Figure 12 illustrates the whole process of array CGH. Generally the used microarray consists of small segments of DNA (oligonucleotides or bacterial artificial chromosomes), which are the targets of the analysis, spotted on a solid support (most of the time a glass slide) [Theisen (2008)]. For the analysis the extracted sample DNA is labeled with a fluorescent dye (in this case producing green light emission) and the control DNA is labeled with another fluorescent dye (in this case red light emission) [Hupe *et al.* (2004)]. The two labeled DNAs are then mixed together and are hybridized to the array (shown in figure 12, step 4). This is possible because the DNAs are denatured and consequently present as single strands [Theisen (2008)]. There should exist two copies of each genomic region of the normal reference DNA in contrast to tumor DNA with gain and losses showing an abnormal number of copies [Hupe *et al.* (2004)]. Afterwards the signal intensities of the hybridized sample and reference DNA are measured and the fluorescent ratio determined. Via computer software it is possible to generate plots to illustrate the lost and gained regions in the sample [Theisen (2008); Hupe *et al.* (2004)].



Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures the fluorescent signals.

Step 6 Computer software analyzes the data and generates a plot.

Figure 12: The whole process of array CGH [Theisen (2008)]

A big disadvantage of array CGH is given by the use of whole genomic DNA purified from tumor tissue because genomically normal cells are nearly always present and these ‘nontumor’ components can dilute the CGH signals. Furthermore with this method only gains and losses can be detected but it is not possible to detect balanced chromosomal aberrations like translocations or inversions [Baslan *et al.* (2012); Charité (2014)]. Hence, methods like next-generation sequencing are of great importance to obviate such shortcomings [Baslan *et al.* (2012)].

1.3.1.2 Next generation sequencing

Sequencing in general plays an important role in the genetic analysis of human disease and is used to resolve basic questions in cellular biology [Shendure, Aiden (2012)]. ‘Next generation sequencing’ represents a group of different high throughput sequencing technologies used for different applications like i) full-genome re-sequencing, ii) mapping of structural rearrangements including for example copy number variations, chromosomal inversions and balanced translocation breakpoints, iii) ‘RNA’-Seq, iv) ‘ChIP’-Seq or v) large-scale analysis of DNA methylation. The greatest advantages of the new next generation sequencing approaches are the possibility to sequence the whole genome and to reduce the costs resulting in an acceleration of research. But all these approaches have one big challenge: to downstream the data management. In general the workflows of the different next generation sequencing approaches are conceptually similar. At first a library preparation with randomly fragmented DNA has to be done. In this process common adapter sequences ligate in vitro to this DNA. Bridge PCR and emulsion PCR are amongst others two common possibilities to generate clonally clustered amplicons that then are sequenced. In the chapter 1.3.1.2.2 and 1.3.1.2.3 two next generation sequencing platforms, their relative strengths, how they work and emerging applications are described [Shendure, Ji (2008)].

1.3.1.2.1 High- throughput Sanger Sequencing

The majority of DNA sequencing production is based on some version of the Sanger biochemistry. There are two possibilities to prepare DNA to be sequenced by high-throughput Sanger sequencing: i) shotgun the novo sequencing - here the randomly fragmented DNA is cloned into a high-copy-number plasmid and transformed in Escherichia coli or ii) via PCR with a primers flanked target. In general Sanger sequencing is a cycle-sequencing reaction including cycles of template denaturation, primer annealing and primer extension [Shendure, Ji (2008)].

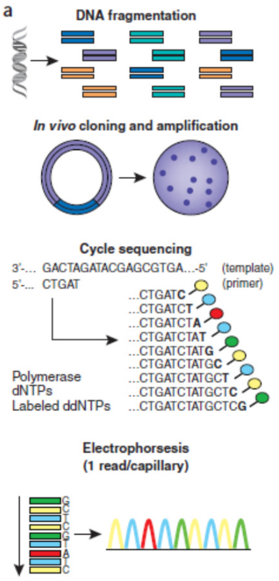


Figure 13: Workflow of Sanger sequencing [Shendure, Ji (2008)]

Figure 13 illustrates the workflow of Sanger sequencing using shotgun libraries. At first the DNA is randomly fragmented, cloned into a high-copy-number plasmid vector and transformed in E coli. Afterwards the DNA of a chosen single colony is isolated for the further processing. Each primers extension process is stopped, if one of the fluorescent labeled dideoxynucleotides is incorporated. This leads to end-labeled DNA strands of different length within which the incorporated dideoxynucleotides represents the nucleotide identity of this terminal position. These obtained products are investigated via high resolution electrophoretic separation in a capillary-based polymer gel. Laser excitation of the fluorescent labels leads to a four-channel emission spectrum which is used to generate the Sanger sequencing trace. Via software this trace is translated into DNA sequences [Shendure, Ji (2008)].

Today the Sanger biochemistry is used to achieve read-lengths of up to approximately 1000 bp and a per-base ‘raw’ accuracy as high as 99.999%. Compared to next generation sequencing, for projects in kilobase- to megabase range, Sanger sequencing is still the technology of choice because of its greater ability ‘to efficiently operate at either small or large production scales’ [Shendure, Ji (2008)].

1.3.1.2.2 454 pyrosequencing

For the 454 system, which was the first available next-generation sequencing product, the clonal sequencing features are generated by emulsion PCR [Shendure, Ji (2008)]. Next to the 454 platform, also the Ion Torrent PGM, the Polonator and SOLiD platforms rely on a kind of emulsion PCR [Loman *et al.* (2012); Shendure, Ji (2008); Quail *et al.* (2012)]. Like shown in figure 14 a PCR amplification of an in vitro-constructed adaptor flanked shotgun library takes place using a water-in-oil emulsion. This method also uses micron-scale beads where one of the PCR primers is fixed. The PCR amplicons are connected to the bead’s surface to the template molecule in emulsion. After the emulsion is broken the beads with the connected amplification products can be selectively enriched. These beads (28 µm) generated by emulsion PCR are used as sequencing features and are randomly added to an array of picoliter-scale wells. In each cycle of pyrosequencing a single nucleotide species is introduced and then a substrate (luciferin) is added. This luciferin is used to generate light at wells where nucleotides are incorporated. Afterwards a washing step is necessary to remove the unincorporated nucleotides [Shendure, Ji (2008)].

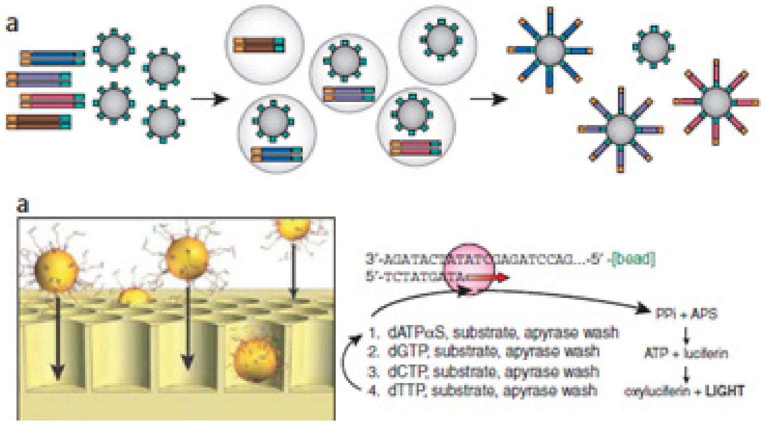

a
a
3'-AGATACTATATCGAGATCCAG...-5' -[bead]
5'-TCTATGATA
1. dATPαS, substrate, apyrase wash
2. dGTP, substrate, apyrase wash
3. dCTP, substrate, apyrase wash
4. dTTP, substrate, apyrase wash
PPi + APS
ATP + luciferin
oxyluciferin + LIGHT

Figure 14: Emulsion PCR [Shendure, Ji (2008)]

The 454 method has a major limitation related to homopolymers, which are consecutive repeats of the same base for example TTT or CCC. Because in the case of multiple consecutive incorporations the length of all homopolymers has to be interfered from the signal intensity [Shendure, Ji (2008)].

For certain applications, for example de-novo assembly, where long read-lengths are critical the 454 instruments may be the method of choice [Shendure, Ji (2008)].

1.3.1.2.3 Illumina genome analyzer

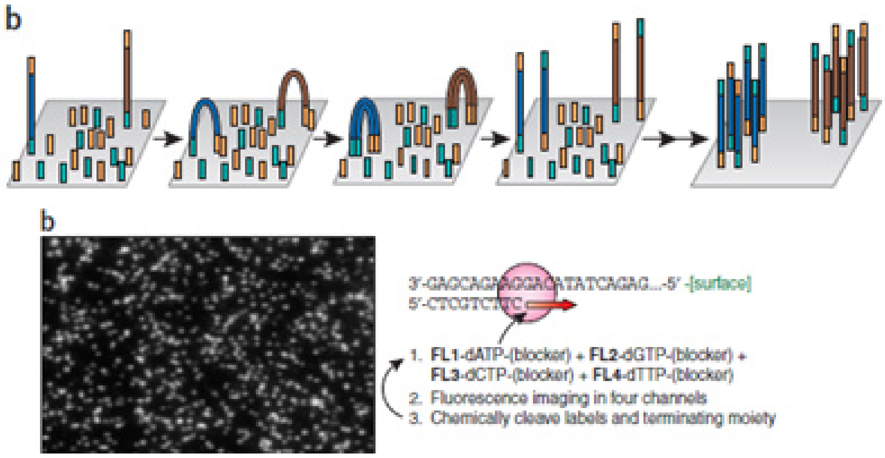

b
b
3'-GAGCAGAAGGACATATCAGAG...-5' -[surface]
5'-CTCGTCTTC
1. FL1-dATP-(blocker) + FL2-dGTP-(blocker) +
FL3-dCTP-(blocker) + FL4-dTTP-(blocker)
2. Fluorescence imaging in four channels
3. Chemically cleave labels and terminating moiety

Figure 15: Bridge PCR [Shendure, Ji (2008)]

The sequencing technology (sequencing by synthesis) of Illumina is the most applied NGS technology worldwide. This technology uses bridge PCR to amplify clonal sequencing features. In the first step a PCR amplification of an in vitro- constructed adaptor flanked shotgun library takes place [Shendure, Ji (2008)]. The used flow cell is covered by a lawn of primers and the single- stranded DNA attaches to the surface of the flow cell channels (figure 15). The solid-phase bridge amplification is initiated by adding unlabeled nucleotides and enzymes. By the incorporation of nucleotides double-stranded bridges are built and a denaturation leads to single-stranded templates which are anchored to the substrate [Illumina (2010)]. Several millions clusters can be generated in each channel of the flow cell in which every cluster can consists of about 1000 clonal amplicons [Shendure, Ji (2008); Illumina (2010)]. Into each sequencing cycle primers, DNA polymerase and four labeled reversible terminators are added. The added polymerase leads to an extension of the sequencing features and after laser excitation the emitted fluorescent from each cluster can be measured and the current base can be identified. After fluorescent imaging a chemically cleavage of the labels and terminating moiety happens [Shendure, Ji (2008); Illumina (2010)]. To determine the sequence of bases in a fragment, the sequencing cycles are repeated. At last an alignment and the comparison of the measured data to a reference take place to identify sequencing differences [Illumina (2010)].

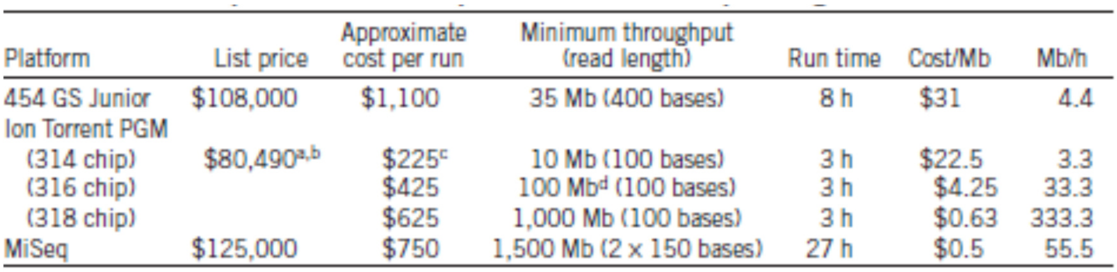
A disadvantage of this method is the limitation of the reading length by factors that cause signal decay and dephasing like for example incomplete cleavage of the label or moiety [Shendure, Ji (2008)].

1.3.1.2.4 Comparison of three high- throughput sequencing platforms

To reconstruct genome sequences from sequence data four major factors are important: Evenness of coverage (coverage is 0 if bases in the genome were not covered by any reads [Quail *et al.* (2012)]), read length, genome depth and read quality. To see which existing approach is the best, Loman *et al.* (2012) compared in their work benchtop high- throughput sequencing platforms (454 GS Junior(Roche), MiSeq (Illumina) and Ion Torrent PGM (Life Technologies) and came to following results [Loman *et al.* (2012)]:

As mentioned before both 454 GS Junior and Ion Torrent PGM are based on emulsion PCR whereas Illumina MiSeq is based on Solexa sequencing-by-synthesis chemistry. Compared to the Illumina HiSeq, MiSeq has a dramatically reduced run time because of a smaller flow cell, reduced imaging time and faster microfluidics [Loman *et al.* (2012)]. However llumina HiSeq is the standard for high throughput massively parallel sequencing whereas the MiSeq system is a lower throughput instrument used in the clinical diagnostic market and smaller laboratories [Quail *et al.* (2012)].

According to the results, MiSeq had the lowest error rates and the highest throughput per run (1.6 Gb/run, 60 Mb/h) but was the longest-running instrument. The 454 GS Junior had the lowest throughput (70 Mb/run, 9 Mb/h) but generates the longest reads (up to 600 bases) and most adjacent assemblies. The Ion Torrent PGM had the fastest throughput (80-100 Mb/h), the shortest run time but produced the shortest reads and had the highest error rate in homopolymeric tracts [Loman *et al.* (2012)].



Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	$108,000	$1,100	35 Mb (400 bases)	8 h	$31	4.4
Ion Torrent PGM						
(314 chip)	$80,490[a,b]	$225[c]	10 Mb (100 bases)	3 h	$22.5	3.3
(316 chip)		$425	100 Mb[d] (100 bases)	3 h	$4.25	33.3
(318 chip)		$625	1,000 Mb (100 bases)	3 h	$0.63	333.3
MiSeq	$125,000	$750	1,500 Mb (2 x 150 bases)	27 h	$0.5	55.5

Figure 16: Comparison of three sequencing platforms [Loman *et al.* (2012)]

Other important factors to be mentioned are the running cost, the speed, set-up and the simplicity of work. Figure 16 shows a list of the cost and efficiency of each instrument. As listed, the Ion Torrent PGM is the lowest-price instrument and the cost per run is the highest for the 454 GS Junior. MiSeq is the highest-price instrument compared to the other two platforms, but has the advantage of the fewest manual steps [Loman *et al.* (2012)].

To draw a conclusion each technology generated useful draft genome sequences [Loman *et al.* (2012)] but each instrument has different advantages and the decision of application should be based on the given task.

1.4 Lung

1.4.1 Structure and function

The lung is an essential respiration organ in the human body and is separated into two lungs. The right lung is divided into three lobes and the left one is divided into two lobes by the fissurae interlobares and each lung is covered by the pleura visceralis. The outside surface of the lung touches the chest wall and the lung is bordered downward by the diaphragm. Between the two lungs the organs of the mediastinum are located. The inhaled air passes the trachea, which divides at the 5th thoracic vertebra into two main bronchi. These main bronchi enter the left and the right lung like shown in figure 17. The main bronchi divide into smaller bronchi, those divide into bronchioles and those again divide into terminal bronchioles. At the 20th division the ductuli alveolares are located which are tightly beset by alveoli [Mutschler *et al.* (2007)]. According to estimations both lungs contain about 300 million alveoli [Mörike *et al.* (1989)]. These alveoli play an important role in gas exchange. By inspiration air with high oxygen content gets in the alveolar space while by expiration air with carbon dioxide is emitted to the environment [Mutschler *et al.* (2007)]. By inspiration the pressure of oxygen in the alveoli is higher than in the venous blood and that is the reason why the oxygen is transported from the alveoli into the blood by diffusion. Otherwise the pressure of carbon dioxide in the venous blood is higher than in the alveoli and so carbon dioxide diffuse out of the blood into the alveoli [Mörike *et al.* (1989)].

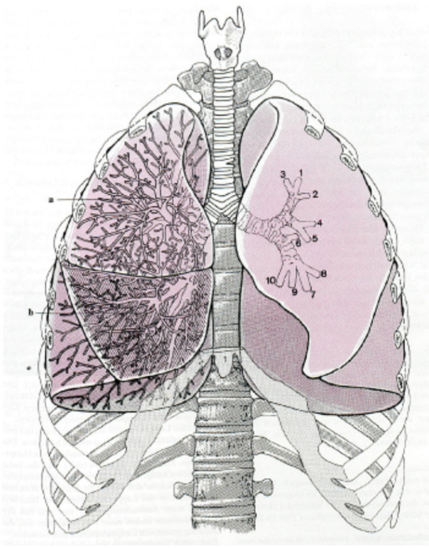


Figure 17: Structure of the lung [Mörike *et al.* (1989)]

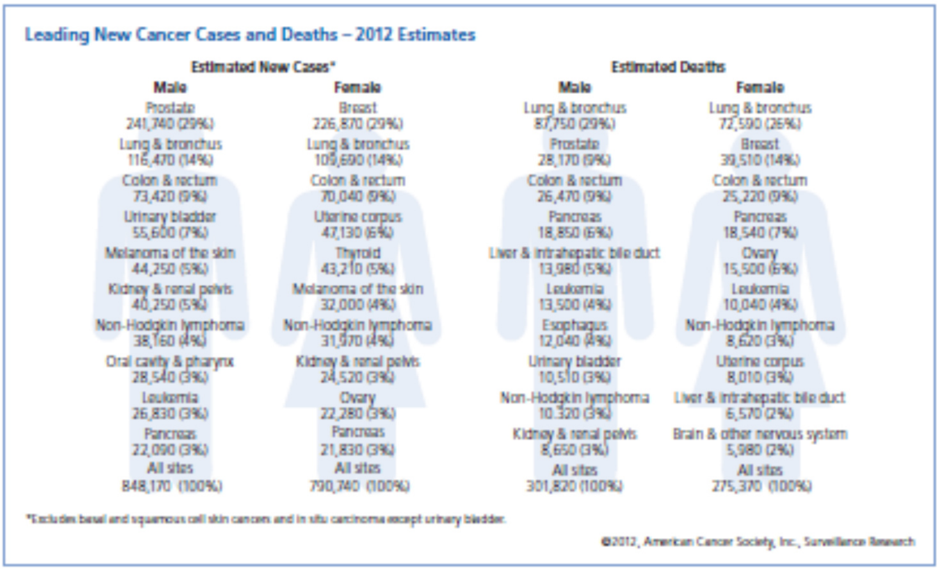
Diseases of the lung are for example different kinds of inflammation like pneumonia or bronchitis, tuberculosis, and bronchial carcinoma [Mörike *et al.* (1989)].

1.4.2 Statistic of cancer

In economically developed countries cancer is the leading cause of death and it is the second leading cause of death in developing countries. The burden of cancer is increasing as a result of population aging and growth and also factors like smoking, lifestyle and diet cause a higher rate of cancer [Jemal et al. (2011)].

According to estimations, in 2008 about 12.7 million cancer cases and 7.6 million cancer deaths have occurred worldwide [Jemal et al. (2011)].

As shown in figure 18 breast cancer in females and prostate cancer in males are the most frequently diagnosed cancer in 2012 but lung and bronchus cancer are the leading cause of cancer death for each sex. Lung cancer in males comprises 29% of the total cancer deaths and 26% in females [Cancer Facts and Figures 2012)].



Leading New Cancer Cases and Deaths – 2012 Estimates

Estimated New Cases*		Estimated Deaths	
Male	Female	Male	Female
Prostate 241,740 (29%)	Breast 226,870 (29%)	Lung & bronchus 87,750 (29%)	Lung & bronchus 72,590 (26%)
Lung & bronchus 116,470 (14%)	Lung & bronchus 109,690 (14%)	Prostate 28,170 (9%)	Breast 39,510 (14%)
Colon & rectum 73,420 (9%)	Colon & rectum 70,040 (9%)	Colon & rectum 26,470 (9%)	Colon & rectum 25,220 (9%)
Urinary bladder 55,600 (7%)	Uterine corpus 47,130 (6%)	Pancreas 18,850 (6%)	Pancreas 18,540 (7%)
Melanoma of the skin 44,250 (5%)	Thyroid 43,210 (5%)	Liver & intrahepatic bile duct 13,980 (5%)	Ovary 15,500 (6%)
Kidney & renal pelvis 40,250 (5%)	Melanoma of the skin 32,000 (4%)	Leukemia 13,500 (4%)	Leukemia 10,040 (4%)
Non-Hodgkin lymphoma 38,160 (4%)	Non-Hodgkin lymphoma 31,970 (4%)	Esophagus 12,040 (4%)	Non-Hodgkin lymphoma 8,620 (3%)
Oral cavity & pharynx 28,540 (3%)	Kidney & renal pelvis 24,520 (3%)	Urinary bladder 10,510 (3%)	Uterine corpus 8,010 (3%)
Leukemia 26,830 (3%)	Ovary 22,280 (3%)	Non-Hodgkin lymphoma 10.320 (3%)	Liver & intrahepatic bile duct 6,570 (2%)
Pancreas 22,090 (3%)	Pancreas 21,830 (3%)	Kidney & renal pelvis 8,650 (3%)	Brain & other nervous system 5,980 (2%)
All sites 848,170 (100%)	All sites 790,740 (100%)	All sites 301,820 (100%)	All sites 275,370 (100%)

*Excludes basal and squamous cell skin cancers and in situ carcinoma except urinary bladder.

©2012, American Cancer Society, Inc., Surveillance Research

Figure 18: Cancer cases and death 2012 [Cancer Facts and Figures 2012)]

Lung cancer, the leading cause of cancer deaths for each sex, will be described more detailed in following chapter.

1.4.3 Bronchial carcinoma

1.4.3.1 Facts and statistics

Lung and bronchus cancer are the leading cause of cancer death and the second frequently diagnosed cancer for both sex in 2012 (see figure 18) [Cancer Facts and Figures (2012)]. 7.6 million deaths in 2008 worldwide resulted from cancer and primary lung cancer accounted for 1.37 million deaths alone. In the United States about 226.160 new cases of lung cancer were reported in 2012 [Xiang *et al.* (2013)].

In Austria the incidence and mortality for men fall since 1986 by over 20% while the incidence and mortality for women are increased by about 10% like shown in figure 19 [Statistik Austria (2013)]. This fact is strongly linked to the factor that in the last decades more women started smoking [Bast Jr, Robert C. *et al.* (2000)].

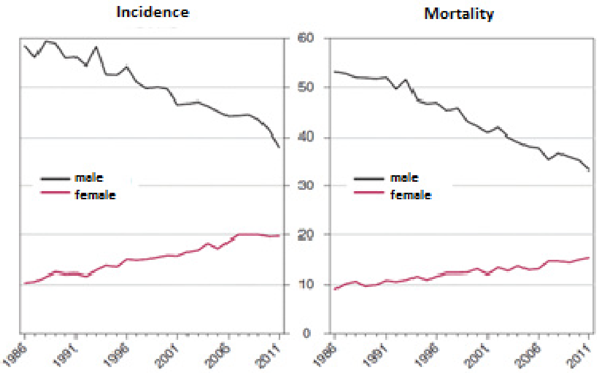


Figure 19: Incidence and mortality of lung cancer patients. Graphic modified from Statistik Austria [Statistik Austria (2013)]

The most important risk factor of bronchial carcinoma is smoking and the risk increases with quantity and duration of smoking [Cancer Facts and Figures (2012); Bast Jr, Robert C. *et al.* (2000)]. The risks of dying of lung cancer are higher for smokers than for non-smokers: male smokers have a 22 times higher risk and female smokers have a 12 time higher risk to die of lung cancer than people who have never smoked [Bast Jr, Robert C. *et al.* (2000)]. The annual incidence rates of lung cancer might be 0.5%, 0.1% and under 0.01% after 45, 30 and 15 years of cigarette smoking [Bast Jr, Robert C. *et al.* (2000)]. This fact shows that the incidence of bronchial carcinoma could be decreased dramatically by stop smoking. Other risk factors are for example radon gas, exposure to secondhand smoke, certain metals like arsenic, asbestos, and radiation [Cancer Facts and Figures (2012)]. Secondhand smoke for example raises the risk of dying of lung cancer by 30%. As mentioned before the risk to get lung cancer after 30 years of smoking is about 0.1%. If a smoker stops after 30 years smoking, the annual risk after 15 further years (in sum 45 years) might be 0.1% instead of 0.5%. Consequently approximately 80% of the risk can be avoided when stop smoking [Bast Jr, Robert C. *et al.* (2000)].

Parts of the lung such as the bronchioles or alveoli as well as the cells lining the bronchi can be the starting point of lung cancer. Lung cancer tends to metastasize before it can be detected on an imaging test (for example x-ray) and that is the reason why it is such a life-threatening disease [American Cancer Society (2014)]. Examples of symptoms of bronchial carcinoma are voice change, persistent cough, chest pain, sputum streaked with blood, and recurrent pneumonia or bronchitis [Cancer Facts and Figures (2012)].

There exist different screenings for detection of bronchial carcinoma like for example chest x-ray, computer tomography, sputum cytology, fluorescence endoscopy and low-dose spiral CT. At present, surgical resection of early stage lung cancer is the only potential curative therapy. Therefore to improve the prognosis, lung cancer has to be detected at an early stage [Xiang *et al.* (2013)]. Results from a large clinical trial showed that lung cancer mortality cannot be reduced by annual screening with chest x-ray. Improved results in detecting lung cancer at earlier, more operable stages in high-risk patients are achieved by using techniques like low-dose spiral computed tomography (CT) scans and molecular markers in sputum [Cancer Facts and Figures (2012)].The 5-year survival rate for all stages combined has improved slowly from 12% in 1977 to only 16% in 2007 [Xiang *et al.* (2013)]. For cases detected when the disease is still localized the 5-year survival rate is 52% but the percentage to diagnose lung cancer at this early stage is only 15% .The 5-year survival is different for small cell lung cancer and non-small cancer. For small cell lung cancer the 5-year survival rate is only 6% whereas it is 17% for non-small cell [Cancer Facts and Figures (2012)]. Figure 20 shows the survival for individuals with lung cancer in England. The median survival for these

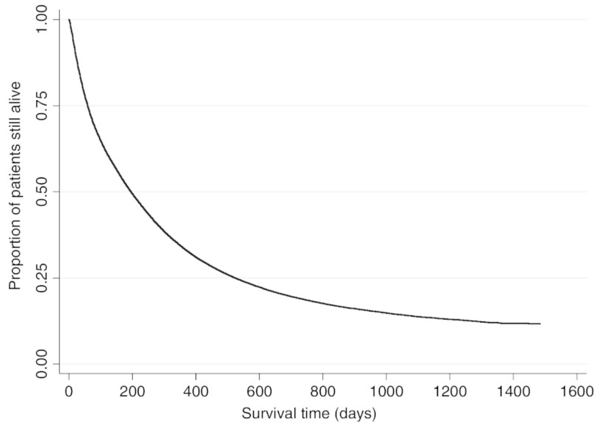


Figure 20: Overall survival of English patients with lung cancer [National Collaborating Centre for Cancer (UK) (2011)]

relevant people is only 203 days and thus less than one year [National Collaborating Centre for Cancer (UK) (2011)].

For purpose of treatment, lung cancer is classified as small cell lung cancer or non-small cell lung cancer. There are different kinds of treatment like surgery, radiation therapy, chemotherapy and targeted therapies whose application depend on type and stage of cancer. Surgery in combination with chemotherapy is usually the treatment of choice for localized non-small cell lung cancers whereas advanced- stage non- small cell lung cancer is treated with chemotherapy and/or targeted drugs. Small cell lung cancer is treated with chemotherapy alone or in combination with radiation [Cancer Facts and Figures (2012)].

1.4.3.2 Two subtypes

Small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) are the two main types of lung cancer [American Cancer Society (2014)] and are described precisely in chapters 1.4.3.2.1 and 1.4.3.2.2. Smoking is the main risk factor of lung carcinoma but is mostly associated with small-cell lung cancer and squamous cell carcinoma. The diagnosis of small-cell lung cancer happens most of the time in an advantage stage and so this disease has a poor prognosis [Herbst *et al.* (2008)].

1.4.3.2.1 Non-small-cell lung cancer

85%-90% of bronchial carcinoma are non-small-cell lung cancer diseases which can be classified into 3 main subtypes: squamous cell carcinoma, adenocarcinoma and large cell carcinoma [American Cancer Society (2014)]. In 2013, german scientists investigated 5000 lung cancer patients in their study and found out that large cell carcinoma as individual subcategory will maybe be needless in the long run because most of the time it can be classified into the other histological subcategories. This detection can change the classification system and further influence the treatment decisions [derStandard (2013)].

Squamous cell carcinoma

Squamous cell carcinomas represent approximately 25% to 30% of all lung cancers. They are often associated with smoking and are found in many cases in the middle of the lungs, near a bronchus [American Cancer Society (2014); Bast Jr, Robert C. *et al.* (2000)]. Generally well- differentiated squamous cell carcinomas grow slowly and have compared to the poorly differentiated type a less probability for distant spread [Bast Jr, Robert C. *et al.* (2000)].

Adenocarcinoma

Adenocarcinoma occurs in current or former smokers and represents about 40% of lung cancer cases [American Cancer Society (2014); Bast Jr, Robert C. *et al.* (2000)]. Nevertheless the most common type of lung cancer in patients who have never smoked is also adenocarcinoma [Herbst *et al.* (2008)]. In comparison to other types, adenocarcinoma is more likely to occur in younger people.

Adenocarcinoma appears usually within the periphery of the lung and it tends to metastasize early to distant sites (e.g. the brain) and regional lymph nodes [American Cancer Society (2014); Bast Jr, Robert C. *et al.* (2000)].

Large cell carcinoma

About 10% to 15% of lung cancers are large-cell carcinomas. This type of lung cancer can appear in any part of the lung and grows and spread most of the time quickly, which complicates the treatment [American Cancer Society (2014);].

1.4.3.2.2 Small-cell lung cancer

This kind of lung cancer accounts for about 10% to 15% and is named after the size of the cancer cells when viewing under a microscope. Smoking as a risk factor is mostly associated with small cell-lung cancer as it is uncommon to have small cell lung cancer as a nonsmoker.
The 'starting- location' of SCLC are often the bronchi near the center of the chest and this small-cell lung cancer tends to spread widely to distant parts of the body before it is found [American Cancer Society (2014)]. SCLC is more prone to develop early metastases than adenocarcinoma, large cell carcinoma and squamous cell carcinoma. Furthermore these three types of non-small cell lung cancer are more frequently amenable to surgical therapy in their early stages than SCLC [Bast Jr, Robert C. *et al.* (2000)].

1.4.3.3 Common genomic variants of lung carcinoma

Lung cancer is a heterogeneous and complex disease and the leading cause of cancer death worldwide [Luo, Lam (2013)].

Activation of cellular oncogenes and inactivation of tumor suppressor genes are critical steps in tumor initiation and progression [Cooper (2000)]. Alterations that play important roles in lung cancer are for example inactivation of the *TP53* tumor suppressor gene [Mechanic *et al.* (2007)] or oncogenic driver mutations like *EGFR* gene mutations, *KRAS* gene mutations, *EML4- ALK* rearrangements and altered *MET* signaling [Luo, Lam (2013)].

In figure 21 a diagram, made from COSMIC (Catalogue of somatic mutations in cancer) is shown including the frequencies of the top 20 genes associated with lung carcinoma. Among other things, the COSMIC database contains somatic mutations which play an important role in cancer development. The information of this database is collected from the literature and sequencing studies carried out by the Cancer Genome Project [Trust Sanger institute (2012a)].

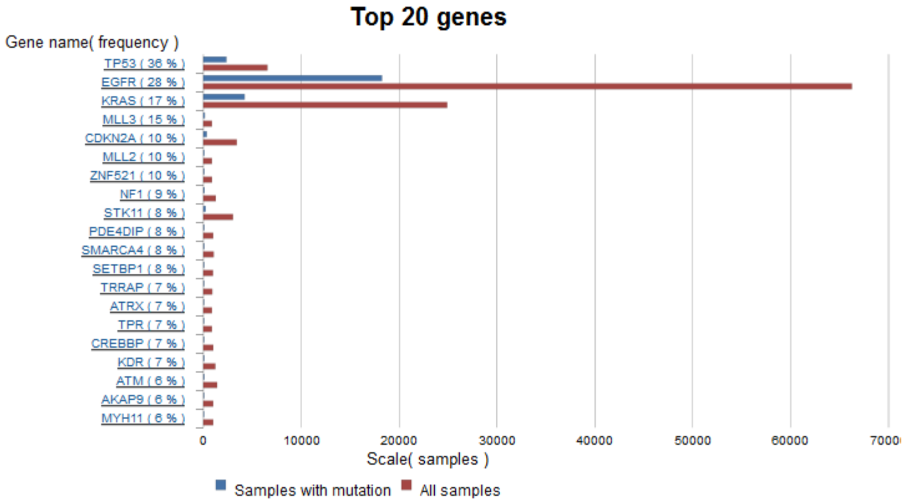


Figure 21: COSMIC diagram of the top 20 genes represented in lung cancer [Trust Sanger institute (2012b)]

1.4.3.3.1 EGFR mutations

The epidermal growth factor receptor is a part of the ErbB family of receptor tyrosine kinases. Next to EGFR the ErbB family of receptor tyrosine kinases contains 3 more receptors: ErbB-2 (HER2), ErbB-3 (HER3) and ErbB-4 (HER4). All four members of the family have a cytoplasmic tyrosine kinase-containing domain, a single hydrophobic transmembrane domain and an external ligand-binding domain [Normanno *et al.* (2006); Bast Jr, Robert C. *et al.* (2000)].

EGFR is a trans-membrane protein and is activated by binding to peptide growth factors of the EGF-family of proteins. EGFR and the other ErbB receptors play an important role in the pathogenesis and progression of different carcinoma types. One example is the induction of cell transformation by EGFR proteins [Normanno *et al.* (2006)].

As shown in figure 22, on average 50% to 70% of lung, colon and breast cancers are related to an expression of EGFR and ErbB-3 [Normanno *et al.* (2006)].

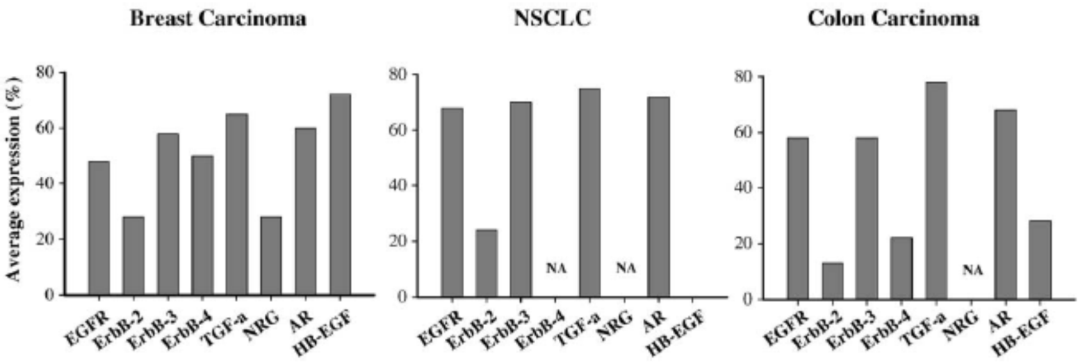


Figure 22: Protein expression of ErbB receptors and cognate ligands in differnt carcinomas [Normanno *et al.* (2006)]

That is one reason why for targeted treatment in NSCLC, *EGFR* gene mutations are the first targets [Luo, Lam (2013)]. In one study, adenocarcinoma showed high levels of erbB2 mRNA whereas erbB2 was not expressed by SCLC cells. A further finding was that erbB2 expression in adenocarcinomas and reduced survival are correlated [Bast Jr, Robert C. *et al.* (2000)].

1.4.3.3.2 KRAS mutations

KRAS, a guanine nucleotide- binding protein, acts as a self- inactivating signal transducer and is one of the most frequently activated oncogenes. An activating mutation in this gene occurs in 17-25% of all human tumors [Kranenburg (2005)].

In NSCLC *KRAS* mutations have an incidence of 8% to 24%. Codons 12 or 13 are the location for the most activating *KRAS* mutations in NSLCL but *KRAS* mutations are also reported in lung adenocarcinoma. A study showed that *KRAS* mutations were found in 15% of tumors from never smokers, 22% of tumors from former smokers and 25% of tumors from current smokers. Another result of this study was that *KRAS* transversion mutations are more common in former/current smokers and *KRAS* transition mutations occur more often in never smokers [Luo, Lam (2013)].

1.4.3.3.3 EML4- ALK rearrangements

'The anaplastic lymphoma kinase (*ALK*) protein is a receptor tyrosine kinase in the insulin receptor superfamily'. According to Soda *et al.* (2007) it is 'a small inversion within the short arm of chromosome 2, resulting in the fusion of the N-terminal of the echinoderm microtubule-associated protein-like 4 (*EML4*) gene with the *ALK* gene' [Luo, Lam (2013)].

Lung adenocarcinoma is the major type showing *EML4- ALK* translocations. *EML4-ALK* fusions are just as *EGFR* mutations associated with never or light smokers [Luo, Lam (2013)].

1.4.3.3.4 MET signaling

According to NCBI 'the proto-oncogene MET product is the hepatocyte growth factor receptor and encodes tyrosine-kinase activity' [NCBI (2014)]. A receptor tyrosine kinase composed of an extracellular α chain and a transmembrane ß-chain is encoded by a *MET* gene which is located on chromosome 7. Many pathways and regulations of various cellular processes like cell proliferation, cell motility, cell invasion et cetera are affected by the activation of *MET* signaling. Due to this fact deregulation of these signaling pathways is a driving force in tumor maintenance and tumor initiation. Overexpression, mutations, alternative splicing or genomic amplification can alter *MET* in transformed cells. For example both *MET* overexpression and *MET* gene copy number variations have been reported in NSCLC. A study including East Asian, Caucasians and African Americans reported that depending on the ethnicity the *MET* mutations can vary. Many *MET*-targeted agents are under clinical and preclinical studies to use *MET* as a major biomarker in lung cancer in future [Luo, Lam (2013)].

1.4.3.3.5 Genetic Variation in *TP53*

TP53 is a tumor suppressor gene and an inactivation of this gene is an early and frequent event in lung carcinogenesis [Mechanic *et al.* (2007)]. The tumor suppressor *TP53* is also called the 'guardian of the genome' because it is able to integrate many signals that are important for the control of life or death. *TP53* gets activated in case of cellular stress, including oncogenic stress and DNA damage and initiates different program, like growth arrest, DNA repair, senescence, and apoptosis, that stop proliferation [Stiewe (2007); Mechanic *et al.* (2007)].

In more than 90% of small cell lung cancer and more than 50% of non-small cell lung cancer mutations in *TP53* are present. P53 protein prevents tumor formation and hence a lung cancer susceptibility can be modulated by genetic variability in the *TP53* gene [Mechanic *et al.* (2007)].

2. Thesis objective

Cancer is the second leading cause of death worldwide and lung cancer is the leading cause of cancer death for males and females. For oncology the analysis of genome copy number variations is of great importance but the use of tumor tissue biopsies do not enable the detection of biomarkers for personalized therapy and only provides a snapshot of mutations at a given time and location [Forshew *et al.* (2012); Heitzer *et al.* (2013c)]. Hence a method that allows repeated analysis of the tumor genomes, for therapies tailored to a patient's individual tumor (personalized medicine), is of great importance. Many studies have shown that there is the possibility for identification of CNV out of cfDNA [Heitzer *et al.* (2013c); Schwarzenbach *et al.* (2011); Leary *et al.* (2012)].

Heitzer *et al.* (2013) established a method called 'plasma-Seq' which is based on Illumimas MiSeq instrument to generate whole genome sequencing of plasma DNA to analyze the tumor genomes noninvasively at low cost within 3 days. This 'liquid biopsy' makes the collection of repeated samples possible and hence changes in cfDNA during cancer treatment and during the natural development of the cancer can be investigated. In their studies, Heitzer *et al.* received useable CNV profiles from this method for patients having breast-, colon- and prostate cancer but till now did not investigate blood samples from lung cancer patients.

Hence, the objective of this thesis was to investigate plasma cfDNA of lung cancer patients based on the method developed by Heitzer et al. to identify specific biomarkers. Blood samples from 5 patients with lung carcinoma and a sum of 6 follow up samples were analyzed. First the Plasma and the cfDNA were extracted and the concentration of cfDNA was measured. Furthermore whole genome sequencing using the Illumina MiSeq instrument and a bioinformatical analysis was performed. To get reliable results it was necessary to apply a number of filtering steps to remove sources of variants. Hence, 20 blood samples from control patients without malignant disease were processed in the same way as the samples from the cancer patients.

The hypothesis underlying to this work is that the complex tumor genomes of patients with lung cancer can be reconstructed out of plasma and CNVs can be found via next generation sequencing. These obtained CNVs of the single patients should be compared and biomarkers for early detection of lung cancer and for the use of personalized medicine should be found.

3. Materials and methods

The figure 23 below shows the general workflow for the whole genome sequencing including isolation of circulating cell free DNA, preparation of a library for sequencing and bioinformatic analysis. The shown steps are explained more detailed in chapters 3.3 to 3.5.

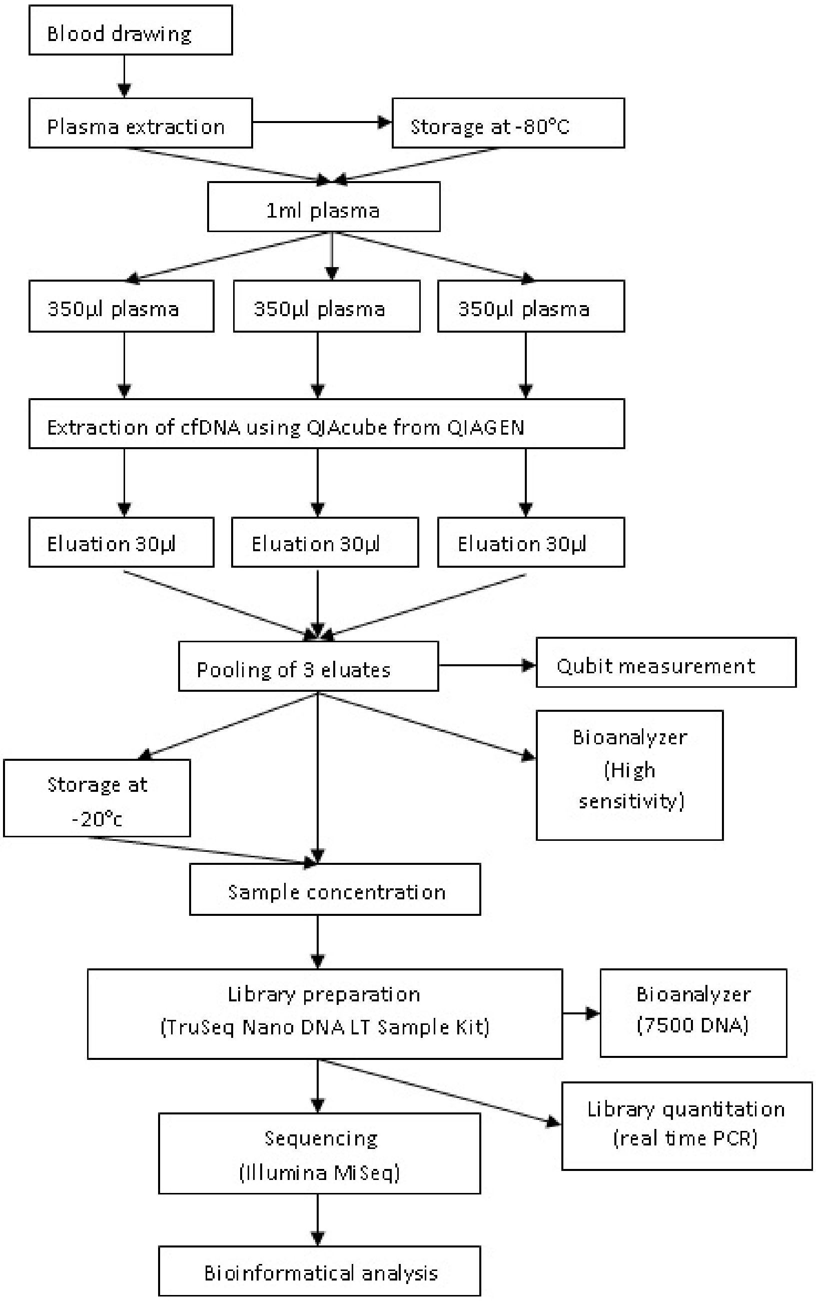


Figure 23: Flow of work for whole genome sequencing

The tables 1 and 2 below show the instruments, software and materials which were used for the work.

Table 1: Used instruments and software

Used Instruments and software	
Instrument	**Company**
Thermomixer compact	Eppendorf
Heraeus™ Fresco 21 Centrifuge	Thermo Scientific
MS3 basic vortexer	IKA®
QIAcube®	QIAGEN
Qubit ® 2.0 Fluorometer	Life Technologies
Centrifuge	Roth®
Syringe Kit	Agilent Technologies
Chip Priming Station	Agilent Technologies
Centrifuge 5417 R	Eppendorf
Agilent 2100 Bioanalyzer + Software *2100 Expert*	Agilent Technologies
Peltier Thermal Cycler	DNA Engine®
Dynal MPC®- M Magnetic Particle Concentrator	Life Technologies
StepOne Plus real time PCR System + StepOne™ software v2.2.2	Applied Biosystems, Life Technologies
Allegra® X- 12R Centrifuge	Beckman Coulter
MiSeq Sequencing System + MiSeq Control Software	Illumina
Pipetman neo	Gilson
Concentrator Plus	Eppendorf

Table 2: Used materials

Used materials		
Material	**Substances of the kit**	**Company**
Buccal Brushes		MasterAmp™
Safe- Lock Tubes 2 ml		Eppendorf
Safe- Lock Tubes 1.5 ml		Eppendorf
Elution Tubes (1.5 ml)		QIAGEN
PCR Soft Tubes 0.2 ml		Bioenzym Scientific
15 ml Tubes		Greiner Bio- One
Qubit ® Assay Tubes		Life Technologies
9 ml K3E K3EDTA Vacuette® Tubes		Greiner Bio-One
Filter- Tips 1000 µl		QIAGEN
Bio Pointe Filter Tips		Bio Pointe Scientific
Tuberculin syringe: 1 ml 25GA x 5/8 in Luer		BD Plastipak™
MicroAmp® Fast Optical 96-Well Reaction Plate with Barcode (0,1 mL)		Applied Biosystems, Life Technologies
MicroAmp® Optical Adhesive Film		Applied Biosystems, Life Technologies
Rotor Adapters		QIAGEN
SYBR® Green		QIAGEN

qPCR Primer (F+R) (10 µM)		Illumina
Stock 1,0 N NaOH		Sigma
NBF (Formalin solution- 10%, neutral buffered)		Sigma- Aldrich
QIAGEN® Proteinase K		QIAGEN
Buccal Swab DNA Extraction Solution		MasterAmp™
Ethanol (100%)		EMSURE®
Nuclease free water or LiChrosolv water		Promega
Dulbecco´s PBS (1x)		PAA
QIAamp® DNA Blood Mini Kit	Buffer AL	QIAGEN
	Buffer AW1	
	Buffer AW2	
	QIAGEN ® Proteinase K	
	QIAamp Mini- spin column	
Qubit ® dsDNA HS Assay Kit	Qubit dsDNA HS Reagent	Life Technologies
	Qubit dsDNA HS Buffer	
	Qubit dsDNA HS Standard #1	
	Qubit dsDNA HS Standard #2	
Agilent High Sensitivity DNA Kit	High Sensitivity DNA Chip	Agilent Technologies
	Electrode Cleaner	
	Spin Filter	
	High Sensitivity DNA Ladder	
	High Sensitivity DNA Marker	
	High Sensitivity DNA Dye Concentrate	
	High Sensitivity DNA Gel Matrix	
Agilent DNA 7500 Kit	DNA Chip	Agilent Technologies
	Electrode Cleaner	
	Spin Filter	
	DNA Ladder	
	DNA Marker	
	DNA Dye Concentrate	
	DNA Gel Matrix	
TruSeq® Nano DNA Sample Preparation Kit (LT)	Resuspension Buffer (RSB)	Illumina
	Sample Purification Beads (SPB)	
	End Repair Mix 2	
	A-Tailing Mix	
	Ligation Mix 2 (LIG2)	
	Stop Ligation Buffer (STL)	
	Enhanced PCR Mix (EPM)	
	PCR Primer Cocktail (PPC)	
	DNA Adapter Indices	
MiSeq® Reagent Kit v3 150 cycles	HT1 (Hybridization Buffer)	Illumina
	Reagent Cartridge	
	PR2 Bottle	
	MiSeq Flow Cell	

3.1 Patients samples

Blood was drawn from five patients with newly diagnosed bronchial carcinoma at the Pulmonology, LKH Graz for later plasma DNA isolation. These patients were neither under medical treatment nor had a surgery to remove parts of the carcinoma at the time of first blood drawing. Patients with small- cell lung carcinoma as well as patients with non-small-cell lung carcinoma were both accepted for this study. Furthermore, a sum of 6 follow up blood samples were taken from these five patients in the course of time.

In figure 24 the frequency of blood drawing for the patient 1 together with chemotherapy cycles and the resulting naming of the samples are shown as an example. Important was to draw the blood before the chemotherapy was set and not during the chemotherapy.

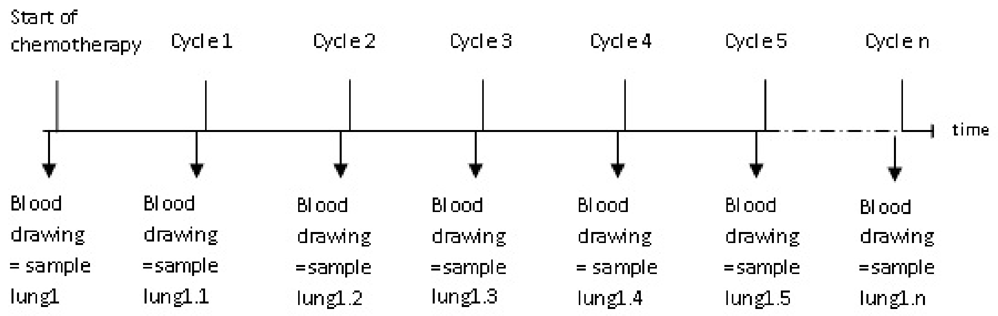

Start of chemotherapy
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Cycle 5
Cycle n
time
Blood drawing = sample lung1
Blood drawing =sample lung1.1
Blood drawing =sample lung1.2
Blood drawing =sample lung1.3
Blood drawing = sample lung1.4
Blood drawing =sample lung1.5
Blood drawing =sample lung1.n

Figure 24: Time table of blood drawing

Additional for each patient a Case Report Form (CRF) was created, which makes the condition of the patient and the condition of the sample at the time of blood drawing respectively reproducible. This CRF is shown in appendix.

To get reliable results it is necessary to apply a number of steps to correctly reduce and validate data [Illumina (2012b)]. For this reason blood was drawn from 11 women (F2- F11, F19) and 9 men (M18-M21, M30- M34) without malignant disease to compare them to the lung cancer patients. These blood samples were processed in the same way as the samples from the cancer patients.

3.2 Isolation of genomic DNA

It is important to isolate genomic DNA (gDNA), to check whether the detected mutation is a somatic mutation (which occurs in a single cell in somatic tissues) or a germinal mutation (which occurs in the germ line and can be passed on to the next generation). Therefore, buccal swab samples from the same patients from which blood was drawn were taken by Pulmonology, LKH Graz.

Table 3: Materials and instruments used for gDNA isolation

Materials and instruments used for gDNA Isolation	
Materials/ instruments	**Company**
Buccal Brushes	MasterAmp™

Buccal Swab DNA Extraction Solution	MasterAmp^{TM}
Thermomixer compact	Eppendorf
Heraeus™ Fresco 21 Centrifuge	Thermo Scientific
MS3 basic vortexer	IKA®

The isolation of genomic DNA from Buccal Swab was performed according to *MasterAmp^{TM} Buccal Swab DNA Extraction Solution* (EPICENTRE) [EPICENTRE]. The author consequently followed the following procedure throughout this thesis research:

At first 500 µl of the MasterAmp^{TM} Buccal Swab DNA Extraction Solution were aliquoted into 1.5 ml tubes. The Buccal Brush was placed into the tube containing DNA Extraction Solution and rotated a minimum of 5 times. Afterwards the brush was pressed against the side of the tube and rotated while removing it from the tube. Then the mix was vortexed for 10 sec and the tube was incubated at 60°C for 30 min. Afterwards the mix was vortexed for 15 sec again and the tube was then transferred to 98°C, incubated for 8 min and vortexed for 15 sec again. Subsequently the tube was returned to 98°C and incubated for an additional 8 min. After further 15 sec of vortexing the tube was briefly chilled on ice to reduce temperature. By centrifugation at 4°C for 5 min the cellular debris were pelleted and the supernatant containing the DNA was carefully transferred to a clean tube. At this point the DNA was stored at -20°C (or at -70°C for a longer term storage).

The obtained genomic DNA was not further processed in the context of this study, but will be investigated in future studies by the Institute of Human Genetics, Graz.

3.3 Isolation and quantitation of plasma DNA

3.3.1 Blood drawing at the clinic

Table 4: Materials and instruments used for blood drawing

Materials and instruments used for blood drawing	
Materials/ instruments	**Company**
Tuberculin syringe: 1ml 25GA x 5/8 in Luer	BD Plastipak™
9 ml K3E K3EDTA Vacuette® Tubes	Greiner Bio- One
NBF (Formalin solution- 10% neutral buffer)	Sigma- Aldrich

This part of the work was conducted by the employees of LKH Graz. Around 9 ml blood was taken in a EDTA Vacuette tube and immediately after blood drawing 225 µl of 10% Neutral buffered Formalin Solution (NBF) was added via tuberculin syringe and mixed by inverting by hand. NBF is used for fixation, prevents cell lysis and preserves cells from decay. This ensures that circulating free DNA and DNA of the cells cannot get mixed and adulterate the results. The blood samples were further processed within 30 min to two hours.

3.3.2 Plasma extraction from whole blood

Table 5: Materials and instruments used for plasma extraction

Materials and instruments used for plasma extraction	
Materials/ instruments	**Company**
15 ml Tube	Greiner Bio- One
Allegra® X-12R Centrifuge	Beckman Coulter
Safe- Lock Tubes 1.5 ml	Eppendorf

In this and the following sections the procedure of plasma and DNA extraction used by the author for this thesis research and the measurements done are explained in all details. At first the volume of EDTA tube was transferred to a 15 ml tube, centrifuged at 200 x *g* for 10 min and afterwards a subsequent centrifugation step at 1600 x *g* for 10 min was done. The supernatant was collected and transferred to a new 15 ml tube and spun at 1600 x *g* for 10min. During the transfer of the supernatant, care must be taken not to transfer leukocytes (white film). After centrifugation, the plasma was carefully transferred to a new 1.5 ml Eppendorf tubes by aliquoting to 1 ml and was stored at -80°C for future use.

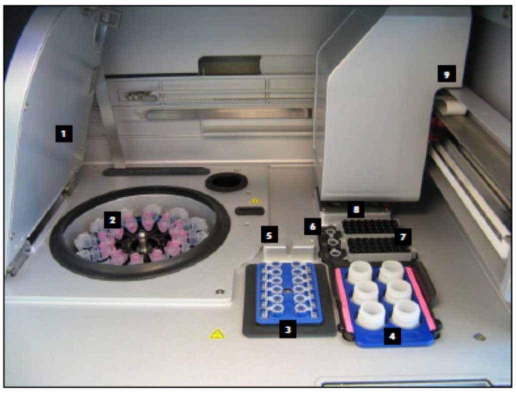
3.3.3 Extraction of free circulating DNA from plasma

The extraction of free circulating DNA from plasma was performed by using the machine QIACube and QIAamp DNA Blood Mini Kit, QIAGEN. The use of QIAcube leads to fully automated spin procedures and so increases the standardization and ease the use. For this procedure QIAamp Mini-spin columns including a silica membrane were used. The DNA was isolated by binding to this membrane and proteins and other contaminants were removed by different washing steps using washing buffers AW1 and AW2. Up to twelve plasma samples can be processed at the same time using QIACube [QIAGEN (2010)].

Table 6: Materials and instruments used for extraction of free circulating DNA from Plasma

Materials and instruments used for extraction of free circulating DNA from Plasma		
Materials/ instruments	**Substances of the kit**	**Company**
QIAamp® DNA Blood Mini Kit	Buffer AL	QIAGEN
	Buffer AW1	
	Buffer AW2	
	QIAGEN ® Proteinase K	
	QIAamp Mini- spin column	
QIAcube®		QIAGEN
Rotor Adapters		
Safe- Lock Tubes 2 ml		Eppendorf
Filter- Tips 1000 µl		QIAGEN
Elution Tubes (1.5 ml)		QIAGEN
Nuclease free water or LiChrosolv water		Promega
Dulbecco´s PBS (1x)		PAA
Ethanol (100%)		EMSURE®

All reagents were prepared according to the protocol *QIAamp DNA Mini and Blood Mini Handbook* and stored at room temperature.



Internal view of the QIAcube.

1 Centrifuge lid
2 Centrifuge
3 Shaker
4 Reagent bottle rack
5 Tip sensor
6 Microcentrifuge tube slots
7 Tip racks
8 Disposal slots for tips and columns
9 Robotic arm

Figure 25: Internal view of the QIAcube [QIAGEN (2008)]

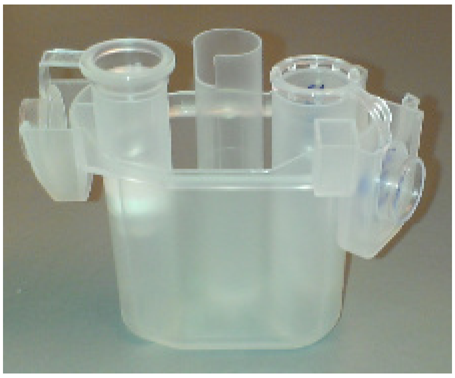


Figure 26: rotor adapter including on the left side the spin column and on the right side the Elution tube

The extraction of free circulating DNA from plasma was performed according to *QIAcube User Manual* (QIAGEN) and *sample & assay technologies* (QIAGEN) [QIAGEN (2008)].

At first the reagent bottle rack was loaded in the following way: The bottles with QIAGEN Buffers (AL, AW1, AW2), nuclease free water and 100% ethanol were filled up and placed on the appropriate position in the reagent bottle rack. Subsequently the bottle rack was inserted in position 4 as shown in figure 25. Afterwards 2 racks with 1000 µl filter- tips were placed in position 8 (see figure 25). Then 1.5 ml Elution tubes and QIAamp Mini- spin column were placed into the appropriate position in each rotor adapter according figure 26. For each plasma sample, three rotor adapters were needed. The loading of centrifuge (position 2 figure 25) depends on the number of samples. Details can be looked up in *QIAcube User Manual* (QIAGEN) Appendix B. Afterwards the samples were loaded and for each plasma sample, three 2 ml Safe- Lock Tubes were needed. The 1 ml plasma was split into the three 2 ml Safe- Lock Tubes a 350 µl. In the case of too less plasma, PBS was used to fill up to 350 µl. The right loading of the shaker rack (position 3 figure 25) can be seen in *QIAcube User Manual* (QIAGEN) Appendix B.

The amount of Proteinase K which was filled in a 2 ml Safe- Lock Tube and placed in position 6 (figure 25) depends on the number of samples. (Details in *sample & assay technologies*; QIAGEN). After the necessary preparations were done the program *QIAamp DNA Blood – blood or body fluid- Plasma DNA extraction* was started.

After approximately 80 min, the program was finished and the 3 x 30 µl elution volume of each sample were pooled together and either stored at -20°C or were used for further analysis.

3.3.4 Qubit 2.0 fluorometer measurement

The Qubit 2.0 fluorometer was used for quantitation of DNA based on fluorescence. By detection of the fluorescence signal, emitted from Molecular Probes® when bounded to specific target molecules, the concentration of DNA could be measured [(Invitrogen 2010a)].

For this part of the work the *Qubit dsDNA HS Assay Kit* was used which is highly selective for double-stranded DNA. This kit is accurate for initial sample concentration from 10 pg/µl to 100 ng/µl [Invitrogen (2010b)].

Table 7: Materials and instruments used for Qubit 2.0 fluorometer measurement

Materials and instruments used for Qubit 2.0 fluorometer measurement		
Materials/ instruments	**Substances of the kit**	**Company**
Qubit ®dsDNA HS Assay Kit	Qubit dsDNA HS Reagent	Life Technologies
	Qubit dsDNA HS Buffer	
	Qubit dsDNA HS Standard #1	
	Qubit dsDNA HS Standard #2	
Qubit® 2.0 Fluorometer		Life Technologies
Qubit ® Assay Tubes		Life Technologies
MS3 basic vortexer		IKA®
Centrifuge		Roth®

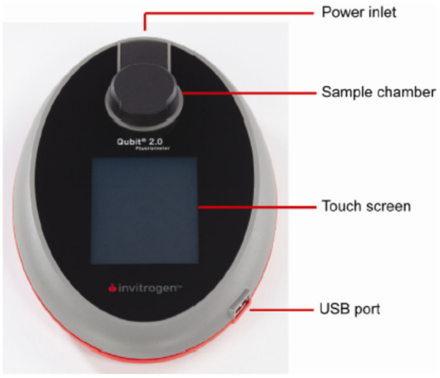


Figure 27: Qubit 2.0 fluorometer [Invitrogen (2010a)]

The measurement happened according to *Qubit dsDNA HS Assay Kits* (Life Technologies) [Invitrogen (2010b)].

For each sample 200 µl Qubit working solution was prepared by diluting the Qubit dsDNA HS reagent 1:200 in Qubit dsDNA HS buffer. Afterwards 190 µl of Qubit working solution were added to each of the tubes used for standards. Later 10 µl of each Qubit standard were mixed to the appropriate tube. For the samples 195 µl of Qubit working solution were added to each sample assay tube and subsequently 5 µl plasma DNA sample were mixed to the 195 µl Qubit working solution (in general sample volume can be varied between 1 µl and 20 µl but the final volume in the tube should be 200 µl). Finally all prepared assay tubes were vortexed and incubated at room temperature for 2 minutes protected from light. After a calibration step with the prepared Standard 1 and Standard 2 mixtures, the sample assay tubes were consecutively placed into the sample chamber (see figure 27) and the concentrations were measured.

3.4 Library preparation

For library preparation the kit *TruSeq® Nano DNA Sample Preparation Kit (low- troughput)* (Illumina) was used. With this kit up to 24 samples can be prepared at one time. During this protocol adapter sequences were added onto the ends of DNA fragments to generate indexed single read or paired-end sequencing libraries [Illumina].

3.4.1 Validation of different library preparation types

To test which of the two existing protocols (*TruSeq® Nano DNA Sample Preparation Kit* or *TruSeq® DNA Sample Preparation Kit*) is more applicable for our experiment, three different libraries were prepared out of plasma obtained from one and the same patient without a malignant disease.

The first library of sample F19.1 was prepared according to the original protocol *TruSeq® Nano DNA Sample Preparation Kit*, the second library of sample F19.2 was prepared according to the original protocol *TruSeq® Nano DNA Sample Preparation Kit*. However the step *Perform End Repair and Size Selection* was omitted and so the library preparation for this sample started at the step *Adenylate 3´Ends*. The step *Perform End Repair and Size Selection* of library preparation of the sample F19.3 happened according to the protocol *TruSeq® DNA Sample Preparation Kit* and the rest (after the first safe stopping point) happened according to the protocol *TruSeq® Nano DNA Sample Preparation Kit*.

The results of the bioanalyzer (shown in chapter 5.2.1) showed that the procedure performed for sample F19.3 was the best for our experiment and is explained in chapter 3.4.4.

3.4.2 Sizing and analysis of DNA fragments via High Sensitivity Bioanalyzer

The *Agilent High Sensitivity DNA Kit* was used for sizing and analysis of DNA fragments. The nucleic acid fragments were separated based on their size via electrophoresis. In general the *Agilent High Sensitivity DNA Kit* can be used for a size range of 50–7000 bp. With this method up to 11 samples can be analyzed at one time [Agilent Technologies (2009)]. This analysis was done only for the samples from cancer patients.

Table 8: Materials and instruments used for sizing and analysis of DNA fragments (high sensitivity Bioanalyzer)

Materials and instruments used for sizing and analysis of DNA fragments		
Materials/ instruments	**Substances of the kit**	**Company**
Agilent High Sensitivity DNA Kit	High Sensitivity DNA Chip	Agilent Technologies
	Electrode Cleaner	
	Spin Filter	
	High Sensitivity DNA Ladder	
	High Sensitivity DNA Marker	
	High Sensitivity DNA Dye Concentrate	
	High Sensitivity DNA Gel Matrix	
Agilent 2100 Bioanalyzer		Agilent Technologies
Chip Priming Station		Agilent Technologies
MS3 vortexer + adapter		IKA®
Syringe Kit		Agilent Technologies
Centrifuge 5417R		Eppendorf
Software *2100 Expert*		Agilent Technologies

Here, procedure followed the *Agilent High Sensitivity DNA Kit Quick Start Guide* (Agilent Technologies) [Agilent Technologies (2009)] and the analysis was done by the software *2100 Expert*.

At first the dilution of the samples for a final amount of 800 pg were calculated based on the Qubit results, to analyze exactly the same amount of each sample to get comparable results.

Preparing the Gel- Dye Mix and setting up the priming station

At first the High Sensitivity DNA dye-concentrate and High Sensitivity DNA gel-mix were brought to room temperature for at least 30 min. Then 15 µl of High Sensitivity DNA dye-concentrate were added to a High Sensitivity DNA gel-matrix-vial, vortexed and spun down. Afterwards this mix was transferred to a spin filter, centrifuged at 2240 g ± 20% for 10min and stored at 4°C or was directly used for the analysis.

The priming station was set up as shown in figure 28 and described in *Agilent High Sensitivity DNA Kit Quick Start Guide*.

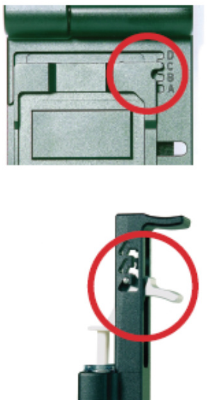


Figure 28: Adjustment of the priming station [Agilent Technologies (2009)]

Loading the DNA chip

At the beginning a new High Sensitivity DNA chip was placed on the chip priming station and 9.0 µl of gel-dye mix were added into the well marked *G* (with black background). The priming station was closed, the plunger was pressed down until it was held by the clip and the clip was released after exactly 60 sec. Subsequently 9.0 µl of gel-dye mix were added into the wells marked *G* (with white background, shown in figure 29). 5 µl of marker were loaded in all sample and ladder wells and afterwards 1 µl of High Sensitivity DNA ladder was added into the marked well. In each of the 11 sample wells 1 µl of sample or additional 1µl of marker (unused wells) was loaded. The chip was placed in the adapter and vortexed for 1 min at 2400 rpm. The chip was then analyzed in the Agilent 2100 Bioanalyzer using the *2100 Expert* software.

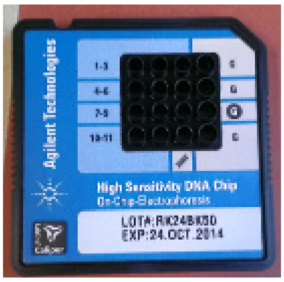


Figure 29: High Sensitivity DNA Chip

3.4.3 Sample concentration

3.4.3.1 Evaporation or dilution of the samples

If the measured concentration obtained with Qubit of the sample was under 10 ng/ml the sample (~85 µl) was evaporated to 50 µl (needed for the end repair step) via speed vac (concentrator Plus from Eppendorf). The mode *V-AQ* (vacuum aqueous) for aqueous solutions was chosen and the concentrator including the opened tubes was run for about 20 min. Samples with a volume under 50 µl after the evaporation step were filled up with nuclease free water to the final volume of 50 µl.

If the concentration of the samples measured with Qubit was more than 10 ng/ml the dilution of the samples were calculated for a final amount of 10 ng according to formula 1. To the calculated amount of sample volume, nuclease free water were added for a final volume of 50 µl.

Formula 1: calculation for samples > 10 ng/ml

$$amount\ of\ \ sample = \frac{10000pg}{\#\ pg\ /\ \mu l\ (Qubit)}$$

3.4.4 Library preparation

To prepare shotgun libraries the kit *TruSeq® Nano DNA Sample Preparation Kit (low- troughput)"* (Illumina) was used.

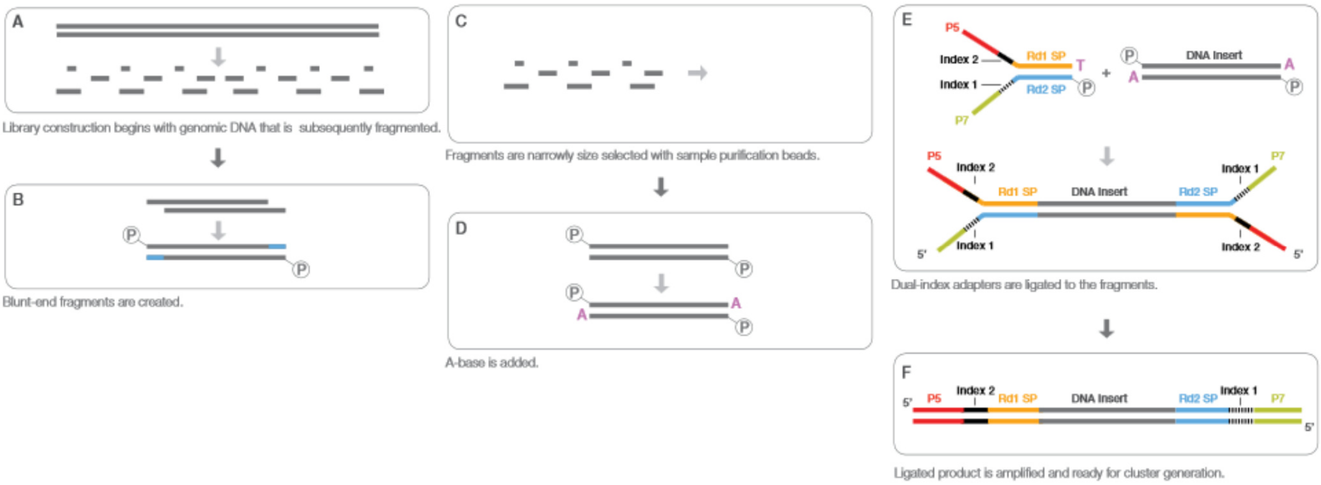
Table 9: Materials and instruments used for library preparation

Materials and instruments used for library preparation		
Materials/ instruments	**Substances of the kit**	**Company**
TruSeq®Nano DNA Sample Preparation Kit (LT)	Resuspension Buffer (RSB)	Illumina
	Sample Purification Beads (SPB)	
	End Repair Mix 2	
	A-Tailing Mix	
	Ligation Mix 2 (LIG2)	
	Stop Ligation Buffer (STL)	
	Enhanced PCR Mix (EPM)	

	PCR Primer Cocktail (PPC)	
	DNA Adapter Indices	
Nuclease free water or LiChrosolv water		Promega
Ethanol (100)		EMSURE®
Safe- Lock Tubes 1.5 ml		Eppendorf
MS3 basic vortexer		IKA®
Centrifuge		Roth®
Peltier Thermal Cycler		DNA Engine®
Dynal MPC®-M Magnetic Particle Concentrator		Life Technologies

The whole library preparation steps were done according to *TruSeq® Nano DNA Sample Preparation Guide (LS)* [Illumina] and *TruSeq® DNA Sample Preparation Guide (LS)* [Illumina (2012a)].

Figure 30 shows the whole workflow of library preparation which will be described more detailed in following chapters.



Library construction begins with genomic DNA that is subsequently fragmented.

Blunt-end fragments are created.

Fragments are narrowly size selected with sample purification beads.

A-base is added.

Dual-index adapters are ligated to the fragments.

Ligated product is amplified and ready for cluster generation.

Figure 30: Workflow of library preparation using the TruSeq® Nano Kit (Illumina) [Illumina (2013a)]

3.4.4.2.1 End repair and size selection

Using this process the overhangs of DNA were converted into blunt ends. The 3' overhangs were removed because of the 3`5`exonuclease activity of the added mix and the 5` overhangs were filled in because of its polymerase activity [Illumina (2012a)].

The step *end repair and size selection* happened as mentioned in 3.4.1 according to *TruSeq® DNA Sample Preparation Guide* (Illumina) [Illumina (2012a)] but reagents from the kit *TruSeq® Nano DNA Sample Preparation Guide* (Illumina) were used.

At first the Resuspension Buffer and the End Repair Mix 2 were thawed at room temperature for 30 min. The thermal cycler was heated to 30°C (lid to 100°C) and 10 µl of Resuspension Buffer were added to the 50 µl of DNA sample. Afterwards 40 µl of End Repair Mix 2 were added to the sample mix. Subsequently the tubes were placed on the pre-heated thermal cycler and were incubated at

30°C for 30 min. During this time the beads were vortexed and a bead mixture for the given amount of samples was diluted: The amount of added Sample Purification Beads was calculated according following equation:

$$Sample\ \ Purificati\ \ on\ Beads\,[\mu l] = \#\, samples\ \ \cdot 160 \cdot 0{,}85$$

The amount of added nuclease free water was calculated according following equation:

$$\text{Nuclease}\ \ \text{free water}\,[\mu l] = \#\, samples\ \ \cdot 160\ \cdot 0{,}15$$

After 30 min on thermal cycler 160 µl of the diluted bead mixture were added to each tube containing 100 µl of End Repair Mix. After mixing, the tubes were incubated at room temperature for 15 min and afterwards placed on the magnetic stand for 15 min. 127.5 µl of the supernatant was removed and discarded twice, 200 µl of freshly prepared 80% EtOH were added to each tube and the mix was incubated at room temperature for 30 sec. Then all of the supernatant was removed and discarded and subsequently the last step was repeated. The opened tubes were placed on the magnetic stand and were dried for 15 min at room temperature. Afterwards the tubes were removed from the magnetic stand and the dried pellet in each tube was resuspended with 17.5 µl Resuspesion Buffer and incubated at room temperature for 2min. The tubes were then placed on the magnetic stand for 5 min and 15 µl of the supernatant were transferred in new tubes. 2.5 µl Resuspension Buffer were added to each tube to reach the required amount of 17.5 µl for the next step. The protocol could safely stopped here (safe stopping point) by storing the samples at -15°C to -20°C.

3.4.4.2.2 Adenylate 3’ends

To prevent the blunt fragments from ligating to one other during the adapter ligation reaction a single ´A´nucleotide was added to the 3´ends of the blunt fragments. For successfully ligation of the adapter to the fragment, the adapter possessed a single ´T´ nucleotide on the 3´end [Illumina].

The procedure of 3’ ends adenylation was performed according to *TruSeq® NanoDNA Sample Preparation Guide* [Illumina].

At the beginning the A-Tailing Mix and Resuspension Buffer were thawed at room temperature. Meanwhile the thermal cycler was pre-programmed with the following program (*ATAIL70*):

- Pre-heat lid to 100°C
- 37°C for 30min
- 70°C for 5min
- 4°C for 5min
- Hold at 4°C

12.5 µl of thawed A-Tailing Mix were added to each sample tube and subsequently the tubes were placed on the pre-programmed thermal cycler and the ATAIL70 program was started. When the thermal cycler temperature had been at 4°C for 5min, the tubes were removed from the thermal cycler and adapters were ligated to the samples (see next chapter).

3.4.4.2.3 Ligate adapters

In this step multiple indexing adapters were ligated to the ends of the DNA fragments [Illumina]. These indices enable the identification of the samples after sequencing a pool of up to 6 samples. The exactly sequences of these adapters are listed in the appendix.

The ligation of the adapters was done according to *TruSeq® Nano DNA Sample Preparation Guide* [Illumina].

At first the Sample Purification Beads, the DNA Adapter Indices and the Stop Ligation Buffer were brought to room temperature for at least 30 min. Meanwhile the thermal cycler was pre-programed with the following program (*LIG*):

- Pre-heat lid to 100°C
- 30°C for 10 min
- Hold at 4°C

Immediately before use the Ligation Mix 2 tube was removed from -15°C to -25°C storage and thawed on ice. At first 2.5 µl of Resuspension Buffer and then 2.5 µl of Ligation Mix2 were added to each tube. Subsequently 2.5 µl of the appropriate thawed DNA Adapter Index were mixed to each tube, the tubes were then placed on the pre-programmed thermal cycler and the *LIG* program was started. After approximately 40 min the tubes were removed from the thermal cycler and 5 µl of Stop Ligation Buffer were mixed to each tube. Then 42.5 µl of well-mixed Purification Beads were added to each tube. Afterwards the tubes were incubated at room temperature for 5min and then placed on the magnetic stand at room temperature for 5 min. 80 µl of the supernatant were removed and discarded from each tube and 200 µl of freshly prepared 80% EtOH were added to each tube. After an incubation time of 30 sec all of the supernatant was removed and discarded. Then the last step was repeated and afterwards the bead pellet was resuspended with 52.5 µl of Resuspension Buffer. The tubes were incubated at room temperature for 2 min and subsequently placed on the magnetic stand for 5 min. 50 µl of the clear supernatant were transferred in new tubes and 50 µl of mixed Sample Purification Beads were added. After an incubation time of 5 min the tubes were placed on the magnetic stand for 5 min again. After this, 95 µl of the supernatant were removed and 200 µl of freshly prepared 80% EtOH were mixed to each tube. After 30 sec of incubation all of the supernatant was removed and discarded. The last step was repeated and afterwards the samples were dried on the magnetic stand at room temperature for 5 min. Afterwards the tubes were removed from the magnetic stand and the dried bead pellet in each tube was resuspended with 27.5

µl of Resuspension Buffer. After 2 min of incubation the tubes were placed on the magnetic stand for 5 min. 25 µl of the clear supernatant were transferred from each tube in a new tube. The protocol could be safely stopped here (safe stopping point) by storing the samples at -15°C to -20°C.

3.4.4.2.4 Enrichment of DNA fragments

To enrich the DNA, fragments have adapter molecules on **both** ends to amplify the amount of DNA in the library using PCR. Therefore a PCR primer cocktail was used that anneals to the ends of the adapters and a sum of 25 PCR cycles were performed. It is important only to amplify DNA fragments having adapter molecules on both ends because later in the sequencing step, fragments without any adapters cannot hybridize to surface-bound primers in the flow cell. In addition fragments having only one adapter can hybridize to flow cell but cannot form clusters [Illumina].

The enrichment of DNA fragments happened according to *TruSeq® Nano DNA Sample Preparation Guide* [Illumina].

At the beginning the PCR Mix and PCR Primer Cocktail were thawed at room temperature and meanwhile the thermal cycler was pre-programmed with the following program (*PCR25*):

- Pre-heat lid to 100°C
- 95°C for 3 min
- 25 cycles of:
  - 98°C for 20 sec
  - 60°C for 15 sec
  - 72°C for 30 sec
- 72°C for 5 min
- Hold at 4°C

At first 5 µl of thawed PCR Primer Cocktail and then 20 µl of thawed Enhanced PCR Mix were added to each tube. Afterwards the tubes were placed on the pre-programmed thermal cycler and the PCR25 program was started. After approximately 1 hour the tubes were removed from the thermal cycler and 50 µl of well mixed Sample Purification Beads were added to each well. After an incubation of 5 min the tubes were placed on the magnetic stand at room temperature for 5 min. Subsequently 95 µl of the supernatant of each tube were removed and discarded and then 200 µl of freshly prepared 80% EtOH were added to each tube. After 30 sec of incubation all of the supernatant was removed and then the last step was repeated. The samples were dried on the magnetic stand at room temperature for 5min and after these 5 min the tubes were removed from the magnetic stand and the dried bead pellet in each tube was resuspended with 32.5 µl of Resuspension Buffer. After 2 min of incubation the tubes were placed on the magnetic stand for 5 min. 30 µl of the clear supernatant were transferred from each tube in a new tube. The protocol could be safely stopped here (safe stopping point) by storing the samples at -15°C to -20°C.

3.4.5 Sizing and analysis of the DNA libraries via DNA 7500 Bioanalyzer

The *Agilent DNA 7500 Kit* was used for sizing and analysis of the prepared DNA libraries. Similar to the High Sensitivity method the nucleic acid fragments were separated based on their size via electrophoresis. The *Agilent DNA 7500 Kit* is in general designed for the sizing and quantitation of double stranded DNA fragments from 100 to 7500 bp. With this method up to 12 samples can be analyzed at one time [Agilent Technologies (2013)].

Table 10: Materials and instruments used for sizing and analysis of prepared DNA libraries (DNA 7500 Bioanalyzer)

Materials and instruments used for sizing and analysis of prepared DNA libraries		
Materials/ instruments	**Substances of the kit**	**Company**
Agilent DNA 7500 Kit	DNA Chip	Agilent Technologies
	Electrode Cleaner	
	Spin Filter	
	DNA Ladder	
	DNA Marker	
	DNA Dye Concentrate	
	DNA Gel Matrix	
Agilent 2100 Bioanalyzer		Agilent Technologies
Chip Priming Station		Agilent Technologies
MS3 vortexer + adapter		IKA®
Syringe Kit		Agilent Technologies
Centrifuge 5417R		Eppendorf
Software *2100 Expert*		Agilent Technologies

The procedure used conformed with *Agilent DNA 7500 Kit Quick Start Guide* (Agilent Technologies) [Agilent Technologies (2009)] and for the analysis the program *2100 Expert was used*.

Preparing the Gel- Dye Mix and setting up the priming station

At first the DNA dye-concentrate and DNA gel-mix were brought to room temperature for at least 30 min. Then 25 µl of DNA dye-concentrate were added to a DNA gel-matrix-vial. Afterwards this mix was transferred to a spin filter, centrifuged at 1500 g ± 20% for 10 min and stored at 4°C or was directly used for the analysis.

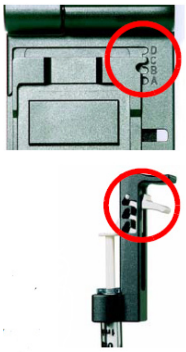


Figure 31: Adjustment of the priming station [Agilent Technologies (2013)]

The priming station was set up as shown in figure 31 and described in *Agilent DNA 7500 Kit Quick Start Guide*.

Loading the DNA Chip

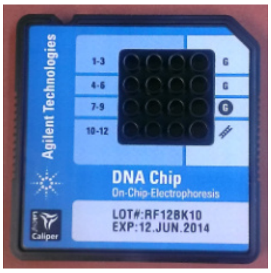


Figure 32: DNA 7500 Chip

At the beginning a new DNA chip was placed on the chip priming station and 9.0 µl of gel-dye mix were added into the well marked *G* (with black background, shown in figure 32). The priming station was closed, the plunger was pressed down until it was held by the clip and the clip was released after exactly 30 sec. Subsequently 9.0 µl of gel-dye mix were added into the wells marked *G* (with white background). 5 µl of marker were loaded in all sample and ladder wells and afterwards 1 µl of DNA ladder was added into the marked well. In each of the 12 sample wells 1 µl of sample or additional 1 µl of marker (unused wells) was loaded. The chip was placed in the adapter and vortexed for 1 min at 2400 rpm. The chip was then analyzed in the Agilent 2100 Bioanalyzer using the *2100 Expert* software.

3.4.6 Quantitation of libraries using qPCR

For next generation sequencing of a DNA library, accurate quantification of the molarity of the DNA library is required. In general library quantification with qPCR is highly sensitive and requires a minimal consumption of sample material. As mentioned in chapter 1.1.4.2 during qPCR the quantity of the targeted DNA library is compared to an already sequenced DNA standard with known concentration.

Table 11 : Materials and instruments used for real time PCR

Materials and instruments used for real time PCR	
Materials/ instruments	**Company**
Nuclease free water or LiChrosolv water	Promega
Standard: Plasma Library 10 nM	
qPCR Primer (F+R) (10 µM)	Illumina
SYBR® Green	QIAGEN
Allegra® X- 12R Centrifuge	Beckman coulter
StepOne Plus real time PCR System + StepOne™ software v2.2.2	Applied Biosystems
MicroAmp® Fast Optical 96-Well Reaction Plate with Barcode (0.1 mL)	Applied Biosystems
MicroAmp® Optical Adhesive Film	Applied Biosystems
PCR Soft Tubes 0.2 ml	Bioenzyme Scientific

At first six dilutions of the *DNA standard* were prepared including: 50 pM; 25 pM; 12 pM; 6.125pM; 3.063 pM and 1.53 pM. Afterwards three dilutions of the *libraries* (e.g. 1:500, 1:1000, 1:2000) depending on the concentration result of the DNA 7500 Bioanalyzer were prepared. Subsequently 7 µl nuclease free water, 1 µl primer, 2 µl library sample and 10 µl SYBR Green were added into each well of a 96 well plate. The reaction of every dilution step (standard and library samples) were run in triplets. Additional a no-template control (NTC), to check the accuracy of the run in retrospect, was included and therefore 2 µl nuclease free water instead of the library sample were added into the

NTC well. The well plate was sealed with a film and centrifuged at 1400 rpm for 2 min. Afterwards the well plate was placed in the instrument and the run was started with following instrument settings:

- Type: Quantification- Standard Curve
- Reagent: SYBR Green
- Speed: fast (40 min)

The used temperature program of qPCR is shown in table 12.

Table 12: Temperature program of qPCR

Program 'Fast'			
	Temperature	Time	Cycles
Stage1	95°C	20 sec	1
Stage2	95°C	3 sec	40
	60°C	30 sec	
Melt Curve	95°C	15 sec	1
	60°C	1 min	
	95°C	15 sec	

Once the qPCR program has finished, the standard curve and the melt curve (exemplarily for 4 samples are shown in figure 33) had to be checked using StepOne™ software. For an accurate measurement the C_T value of all samples (blue squares) should lie exactly on the standard curve.

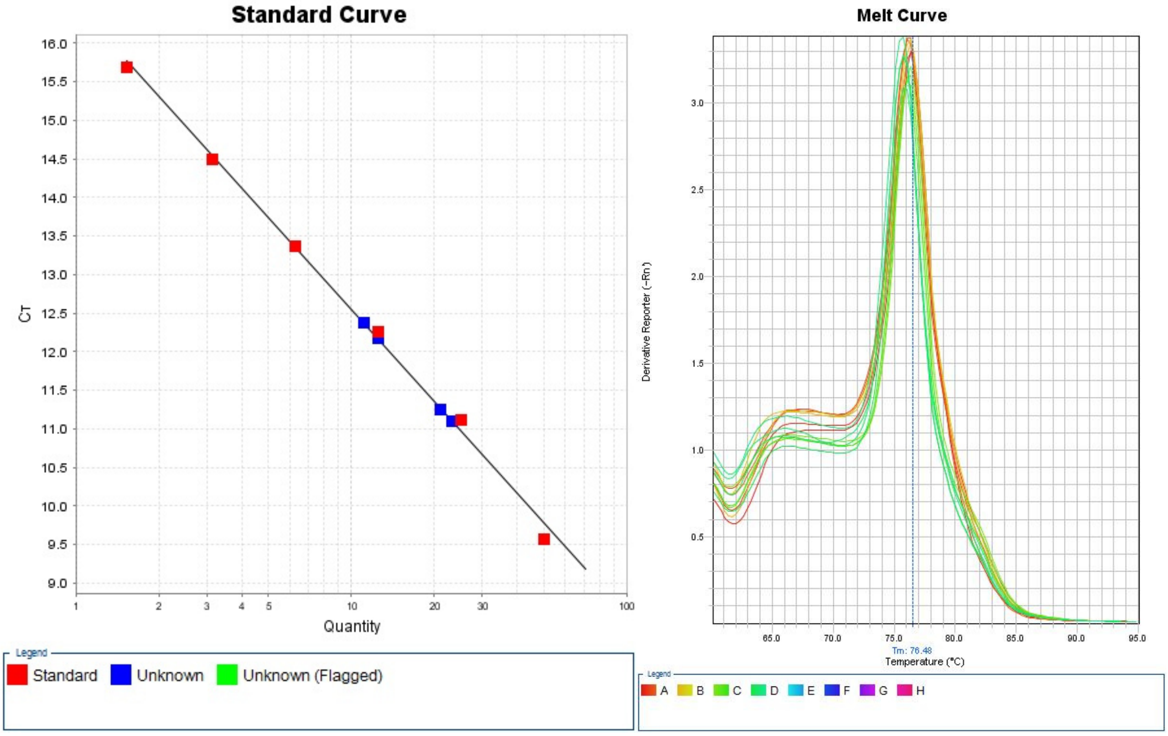
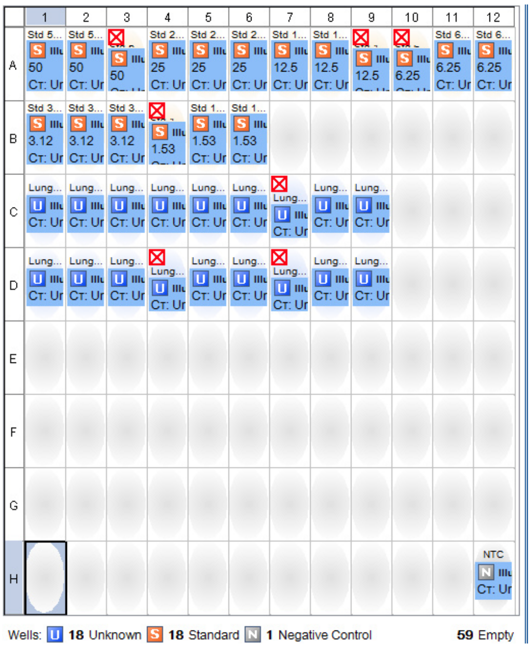


Figure 33: Standard and melt curve of 4 samples analyzed with qPCR

For a better result every sample was put on the 96-well plates as triplicates. Outliers of the amplification plot (exemplarily shown in figure 35 for the samples Lunge3.4 and Lunge4.2) were omitted in the plate layout as shown in figure 34. Afterwards the filtered data were exported in an excel file and the real concentration (nM) of each sample was calculated. Therefore the mean of the quantity of the triplicates was calculated and according to the real concentration of each sample the samples were ready to be diluted to four nM for sequencing.



	1	2	3	4	5	6	7	8	9	10	11	12
A	Std 5... S Illu 50 CT: Ur	Std 5... S Illu 50 CT: Ur	☒ S Illu 50	Std 2... S Illu 25 CT: Ur	Std 2... S Illu 25 CT: Ur	Std 2... S Illu 25 CT: Ur	Std 1... S Illu 12.5 CT: Ur	Std 1... S Illu 12.5 CT: Ur	☒ S Illu 12.5	☒ S Illu 6.25	Std 6... S Illu 6.25 CT: Ur	Std 6... S Illu 6.25 CT: Ur
B	Std 3... S Illu 3.12 CT: Ur	Std 3... S Illu 3.12 CT: Ur	Std 3... S Illu 3.12 CT: Ur	☒ S Illu 1.53	Std 1... S Illu 1.53 CT: Ur	Std 1... S Illu 1.53 CT: Ur						
C	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	☒ Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur			
D	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	☒ Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	☒ Lung... U Illu CT: Ur	Lung... U Illu CT: Ur	Lung... U Illu CT: Ur			
E												
F												
G												
H												NTC N Illu CT: Ur

Wells: U 18 Unknown S 18 Standard N 1 Negative Control 59 Empty

Figure 34: Plate layout for qPCR

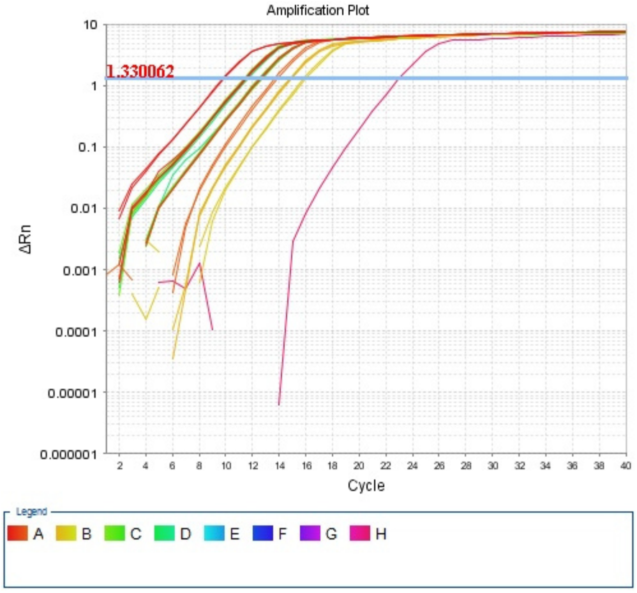


Figure 35: Amplification plot exemplarily for samples *Lunge3.4* and *Lunge4.2*

3.4.7 Sample preparation for sequencing on the MiSeq

For sequencing the libraries on the MiSeq, the libraries had to be denatured and diluted. For whole-genome sequencing, a maximum of 6 plasma samples with different indices had to be pooled equimolarily. The amount of samples was limited to 6, because a sum of 20 million reads were available for sequencing on the MiSeq and each sample needed 3-4 million reads to provide good results.

Table 13: Materials and instruments used for sample preparation for sequencing

Materials and instruments used for sample preparation for sequencing	
Materials/ instruments	**Company**
HT1 (Hybridization Buffer)	Illumina
Stock 1.0N NaOH	Sigma
Nuclease free water	Promega

The preparation of the libraries for sequencing was done according to *Preparing Libraries for Sequencing on the MiSeq* (Illumina) [Illumina (2013)].

At first 1 ml of 0.2N NaOH was prepared by combining 980 µl nuclease free water and 20 µl Stock 1.0N NaOH. Subsequently 4 nM libraries were prepared and 5 µl of each library were pooled together in a new tube (=4 nM sample DNA). To denature the DNA, 5 µl of 4 nM sample DNA and 5 µl of freshly diluted 0.2N NaOH were combined in a new tube and incubated for 5 min at room temperature. Afterwards 990 µl pre-chilled HT1 were added to the tube containing denatured DNA. This resulted in a 20 pM denatured library in 1mM NaOH. The denatured DNA was placed on ice until the final dilution of for example 12 pM could be prepared. Therefore 360 µl of 20 pM denatured DNA and 240 µl of pre-chilled HT1 were combined.

It was important to consider that the denaturation step must be performed only when the Reagent Cartridge (figure 36) was already thawed.

3.5 Sequencing with Illumina MiSeq

The Illumina MiSeq system is based on the Solexa sequencing-by-synthesis chemistry and works similarly like described before in chapter 1.3.1.2.3. Using this instrument whole-genome sequencing is possible and copy number variations can be detected [Heitzer *et al.* (2013c)]. The selected read type was *single read* and an amount of *151* cycles reads were chosen.

Table 14: Materials and instruments used for sequencing with Illumina MiSeq

Materials and instruments used for sequencing with Illumina MiSeq		
Materials/ instruments	**Substances of the kit**	**Company**
MiSeq® Reagent Kit v3 150Cycles	HT1 (Hybridization Buffer)	Illumina

	Reagent Cartridge	
	PR2 Bottle	
	MiSeq Flow cell	
MiSeq Sequencing System + MiSeq Control Software		Illumina

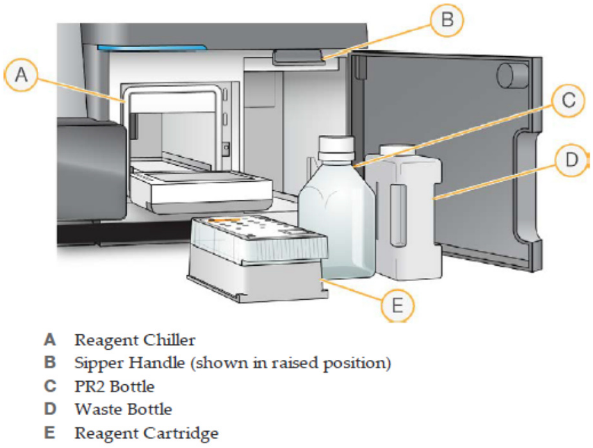


Figure 36: Reagent compartment of MiSeq sequencing system [Illumina (2012b)]

The loading of the MiSeq was done according to *MiSeq System User Guide* (Illumina) [Illumina (2012b)].

At the beginning the Reagent Cartridge (figure 36) was thawed in a water bath at room temperature. The diluted and denatured library mix (600 µl) was then loaded onto the Reagent Cartridge in the designated reservoir. The flow cell was washed and dried and placed on the allocated space. Then the PR2 bottle (figure 36) and the Reagent Cartridge were loaded. Afterwards the run was started, run for approximately 12 hours and subsequently a post-run wash was done.

The analysis of the generated data is discussed in chapter 4.

4 Data evaluation

Sequencing with the MiSeq System generates an enormous amount of data that have to be analyzed to identify aneuploidy of the genomes. Therefore the data were aligned to a standard human genome and steps like filtering and normalization had to be performed as illustrated in following chapters. The procedure based on the paper *Genome- wide copy number analysis of single cells* (Baslan *et al.* (2012)) was developed by the computer scientist Peter Ulz of the Human Genetics Institute, Graz.

4.1 Align against human genome (hg19)

With the aid of a specific algorithm the reads per chromosome arm were counted and afterwards aligned against the human genome (hg19) using the BWA algorithm (Burrows-Wheeler Alignment). This was done in order to find matches to the human genome in the corresponding FASTQ-file where all sequence tags were saved which passed the Illumina chastity filter. The results were saved as an aligned file [Heitzer *et al.* (2013c)]:

```
bwa   aln   -f   'Sample'.aln  -t   20   ~/RefSeq/hg19/hg19.navin   'Fastq_file'
```

Matches founded in this process had to be translated into a text based format which is readable for humans. This was done by converting the aligned file to a SAM-file (Sequence Alignment/Map) in which for example the chromosomal positions of the reads, the sequence, the quality and the mapping quality were saved. The quality values were coded as Phred Scores which are defined as following:

$$\text{Phred Score}= -10\ \log_{10}{}^{\circ}\text{pe}$$

'pe' represents the probability that the mapping position is wrong.

Finally the SAM file was re-converted to a BAM file, because many tools rely on the binary BAM format, which allow, among other things, a faster calculation.

```
samtools   view   -S   -b   -o   'Sample'.bam   'Sample'.sam
```

4.2 Remove PCR duplicates

Samtools rmdup was used to remove PCR duplicates:

```
samtools   rmdup   -s   'Sample'.bam   'Sample'.rmdup.bam
```

This is necessary because limited-cycle PCR was done during library preparation. Fragments with the same start and end coordinates as well as the same sequences were removed in this step. Since the coverage in our analyses was very low, nearly only duplicates were removed.

4.3 Count reads in genomic bins

At first the whole genome was divided in genomic windows/bins. The determined reads were randomly located within these windows and the reads within each genomic window were calculated via a python script.

```
Script.py    'Sample'.sam    'Sample'.bincounts     'Sample'.stats
```

On average a sum of 50.000 genomic windows were created with an average size of 56,344 kbp and one genomic bin contained usually 50-100 reads. [Heitzer *et al.* (2013c)]

4.4 Correction of the GC bias and normalization

Via *LOWESS* function the GC (guanin-cytosin) bias, which is represented in regions with a very high or very low GC-content was corrected.

To avoid position effects as wells as variations at centromeres and telomeres respectively the sequencing data were normalized with GC-normalized read counts of the healthy controls and the $\log_2$ ratio was calculated [Heitzer *et al.* (2013c)].

4.5 Calculation of $\log_2$ ratio and z-score

Furthermore a CGHweb analysis was done. The CGHweb framework in R performs segmentation by using the CBS (Circular Binary Segmentation) algorithm and the GLAD algorithm. Genomic windows with similar copy number were combined and the mean value of the reads included in the segments was determined. Overrepresented regions (amplifications) have a higher mean value in contrast to underrepresented regions with a lower mean value. The $\log_2$ratio can be calculated as following:

$$\log_2 ratio = \frac{corrected\,read\,counts(sample)}{corrected\,read\,counts(control)}$$

A negative $\log_2$ratio represents a deletion whereas a positive $\log_2$ratio represents an amplification. The cutoff in diagnostics normally is ±0.2. The problem that appeared during the analysis for this thesis is the unknown ratio of normal DNA and tumor DNA. By a cutoff of ±0.2 genomic changes may not be detected. This fact makes a z-score analysis for this thesis indispensable.

The segments created via CGHweb framework were also used for calculation of the segmental z-scores. Z- scores were calculated by subtracting the mean read-count ratio of control samples from the read-count ratio of the sample and dividing through the standard deviation read-count ratio of control samples [Heitzer *et al.* (2013c)].

$$Z_{region} = \frac{ratio_{sample} - mean(ratio_{controls})}{SD(ratio_{controls})}$$

The distribution of reads in a genomic window should be an equal distribution with a mean value of 0 and a standard derivation of 1. Is the deviation of the z-score from the mean value higher as the threefold standard derivation of non-cancer controls, it is most likely that in this region a true deletion or amplification exists. A z-score under 3 represents the loss of parts of the chromosome or the whole chromosome whereas a z-score higher than 3 represents gains of parts of the chromosome or the whole chromosome. The outliers of z-scores of each patient are shown in the appendix.

5 Results

To normalize possible artifacts occurring during sequencing additional blood was drawn from 11 women (F2- F11, F19) and 9 men (M18- M21, M30- M34) without malignant disease and these blood samples were processed in the same way as the samples from the cancer patients.

5.1 Qubit 2.0 fluorometer measurement

To quantify the contained DNA the concentration of DNA was measured via Qubit 2.0 fluorometer measurement. The average concentration of the DNA in plasma samples from the cancer patients was 36.4 ng/ml and the average concentration of the DNA in control samples was 4.661 ng/ml. In average the concentration of DNA in the plasma obtained from cancer patients was 9 fold higher than in the control samples. The single results are shown in table 15.

Table 15: Qubit measurement results from samples and controls

Qubit results samples			
Sample	**Concentration (ng/ml Plasma)**	**Control**	**Concentration (ng/ml Plasma)**
Lunge 1	10.72	M18	8.64
Lunge 1.1	15.68	M19	4.45
Lunge 1.2	14.08	M20	3.68
Lunge1.3	12.64	M21	3.62
Lunge 2	127.20	M30	3.94
Lunge 3	24.16	M31	4.48
Lunge3.2	10.64	M32	2.24
Lunge3.4	44.16	M33	3.01
Lunge 4	16.32	M34	4.67
Lunge4.2	84.8	F2	3.71
Lunge5	40	F3	9.60
		F4	2.43
		F5	4.19
		F6	3.94
		F7	3.65
		F8	4.03
		F9	2.75
		F10	4.96
		F11	3.71
		F19	11.52

5.2 Library preparation

5.2.1 Validation of different library preparation types

To test which library preparation protocol is more applicable for this study, three different libraries were prepared out of plasma obtained from one and the same patient without a malignant disease (see chapter 3.4.1). The results of the Bioanalyzer (DNA 7500 Kit) analyses are shown in figure 37.

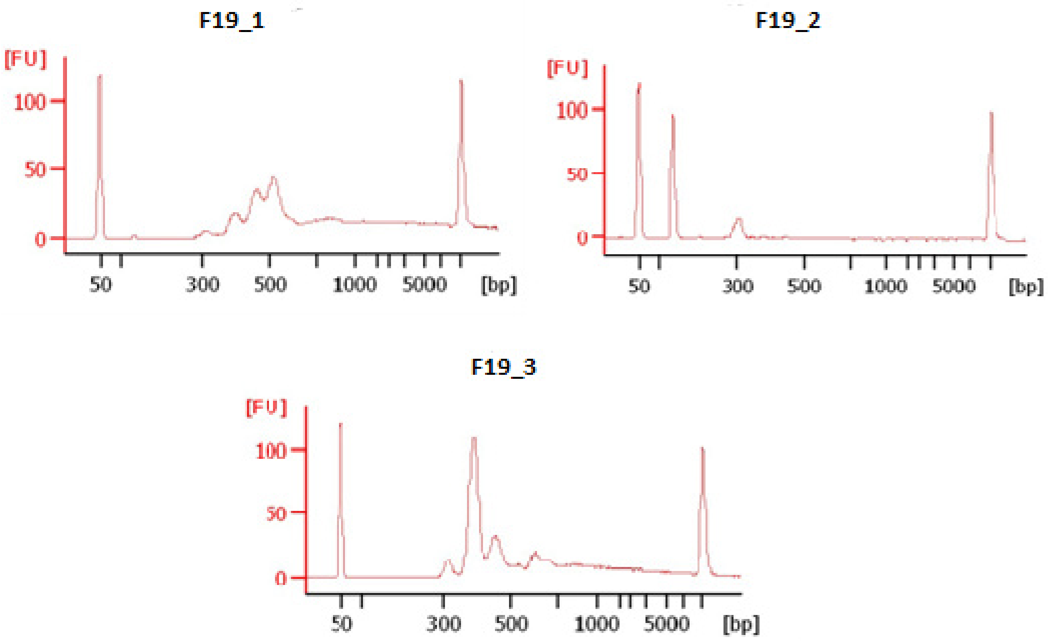


Figure 37: Results of library preparation validation

The result for sample F19_2 showed an adapter peak only at a length of about 120 bp, which is an indication for failed ligation of the adapters to the DNA fragments during library preparation. Both, the samples F19_1 and F19_3 showed an enrichment of DNA fragments with a size between 300- 500 bp but because of the higher enrichment of DNA fragments, the method used for preparing sample F19_3 was chose for further library preparations.

5.2.2 High sensitivity Bioanalyzer

The high sensitivity Bioanalyzer was used to determine the length of DNA fragments via electrophoresis. All samples showed an enrichment of fragments in the range between 150 bp and 280 bp. Figure 38 shows the result from sample *Lunge1* exemplarily for all other results which are shown in appendix. This sample had an enrichment of DNA fragments with an average length of about 170 bp.

In figure 39 the result of a hemolytic blood sample is shown exemplarily. Hemolysis means that the red blood cells (erythrocytes) ruptured and their high molecular content can be released into the plasma. These high molecular DNA fragments were represented by the peaks next to the high 10380 bp peak. These DNA fragments could manipulate the sequencing results and hence the result of this sample should be interpreted with caution.

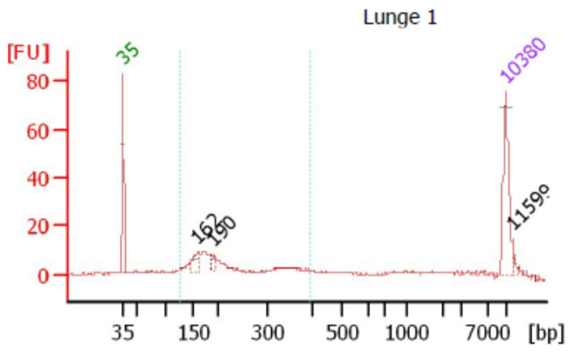


Figure 38: High Sensitivity Bioanalyzer result for sample Lunge1

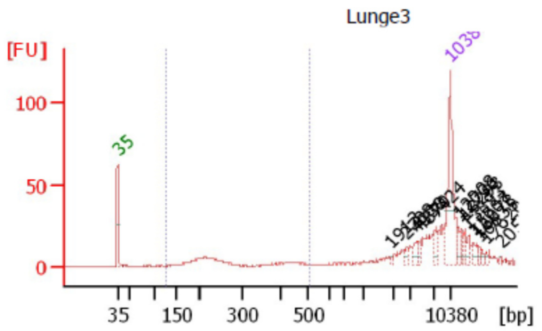


Figure 39: High sensitivity Bioanalyzer result for sample Lunge3

5.2.3 DNA 7500 Bioanalyzer

The high sensitivity Bioanalyzer is used to determine the quality and quantity of the libraries after library preparation and amplification. The range of length of the DNA fragments for the lung samples was in average between 300 and 700 bp. Figure 40 shows the result of the library from sample *Lunge 1* exemplarily for all other results which are shown in appendix.

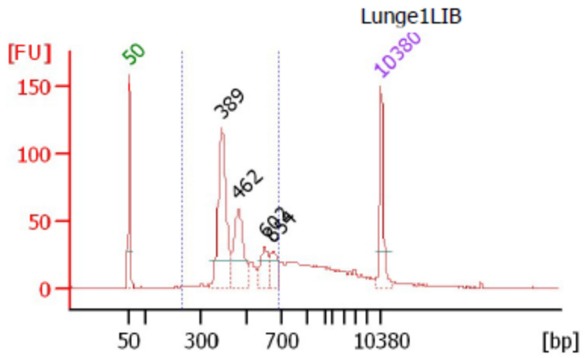


Figure 40: DNA 7500 Bioanalyzer result for sample Lunge1

The concentration determined via DNA 7500 Bioanalyzer (see table 16) was then used to calculate the dilutions series of the samples and controls needed for the real time PCR. According to the Bioanalyzer results the samples were diluted to 20 nM for further processing.

Table 16: Concentrations of samples and controls measured via DNA 7500 Bioanalyzer

Results of DNA 7500 Bioanalyzer			
Sample	**Concentration (nmol/l)**	**Control**	**Concentration (nmol/l)**
Lunge 1	78.1	M18	74.8
Lunge 1.1	103.1	M19	70.2
Lunge 1.2	102.8	M20	90.2
Lunge1.3	72.5	M21	89.9
Lunge 2	95.1	M30	76.6
Lunge 3	80.7	M31	87
Lunge3.2	101.7	M32	85.5
Lunge3.4	80.9	M33	82.8
Lunge 4	77.2	M34	83
Lunge4.2	86.7	F2	67.3
Lunge5	63.9	F3	69.1
		F4	65.5
		F5	60.7
		F6	86.9
		F7	110.5
		F8	118.9
		F9	119.7
		F10	93.4
		F11	87.7
		F19	85.4

5.2.4 Real time PCR

For an accurate determination of the amplified DNA fragments real time PCR was used. Before qPCR analysis was started, the samples were diluted to 20 nM on the basis of the DNA 7500 Bioanalyzer results. Compared to the 20 nM dilutions based on Bioanalyzer results, the concentrations determined using qPCR were most of the time higher or lower than 20 nM. The Bioanalyzer provided different results because it quantifies all DNA-fragments. On the contrary by using qPCR only these fragments with both adaptors ligated to the ends were detected. Hence, the dilution of libraries to 4 nM for sequencing was done according to the real time PCR results and these results are listed in table 17.

Table 17: Results of real time PCR

Real time PCR results			
Control	**Concentration (nM)**	**Sample**	**Concentration (nM)**
M18	25.99	Lunge 1	62.67
M19	24.28	Lunge 1.1	19.56
M20	26.49	Lunge 1.2	18.44
M21	22.32	Lunge1.3	20.11
M30	21.47	Lunge 2	34.04
M31	22.72	Lunge 3	18.98
M32	22.76	Lunge3.2	15.73
M33	28.37	Lunge 4	20.57
M34	20.27	Lunge3.4	24.02
F2	33.44	Lunge4.2	21.83
F3	30.90	Lunge5	9.4
F4	41.36		
F5	46.61		
F6	20.68		
F7	17.40		
F8	18.97		
F9	18.52		
F10	20.41		
F11	17.83		
F19	739.98		

5.3 Sequencing with Illumina MiSeq

In figure 41 the copy number variation profile of sample *Lunge2* is shown. This sample was the only one which showed significant gains and losses and hence it will be described more detailed in this chapter. All other genome profiles, excepted *Lunge5* which will be described below, were practically balanced (see appendix). In general, the increase of the base line at the $\log_2$ ratio of zero corresponds to the gain of this region and significant gains are marked with red. Losses of regions which are marked with blue correspond to decreases of the base line. For the sample *Lunge2* losses in the chromosome arms 8p and 20p were detected and also a visible, but not marked, decrease of the base line appeared in chromosome arm 4q. In addition, in chromosome arms 1q and 7q gains were detected and a visible but not marked increase of the base line appeared in chromosome arm 5p. The most of the other small changes appeared in centromere and telomere regions as for example in chromosomes 3 or 15 and hence, were most likely artifacts caused by the difficult alignment of reads because of the repetitive sequences within these regions. As mentioned before the $\log_2$ ratio was calculated by division of the corrected read counts of the sample by the corrected read count of the control. Since women have no chromosome Y (=chr24) there is a division of 0 by 0 and hence a random result was obtained for chr24 for all women.

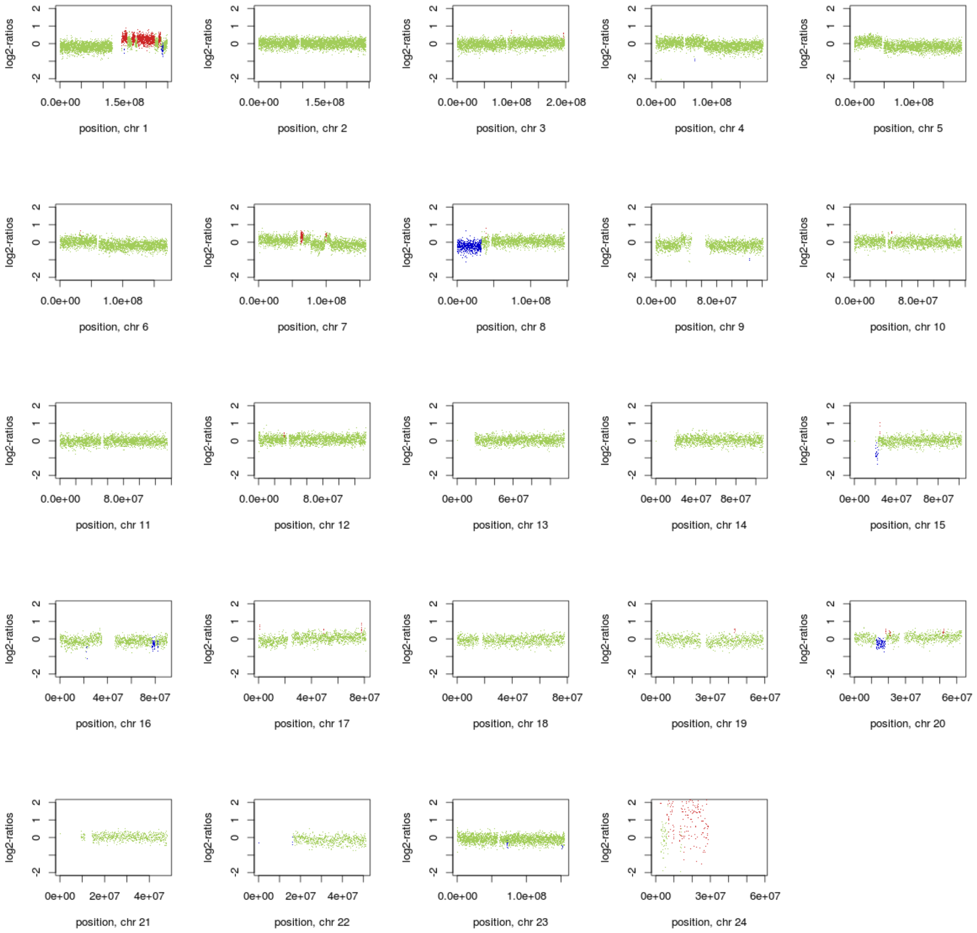


Figure 41: Copy number variation profile for sample *Lunge2*

Figure 42 shows the copy number variation profile for sample *Lunge5*. Based on experience it can be said that probably artifacts during sequencing are the reasons for such a high number of gains and hence the sequencing results (z-scores etc.) of this sample will not be included in the evaluation. An inclusion would maybe distort the results.

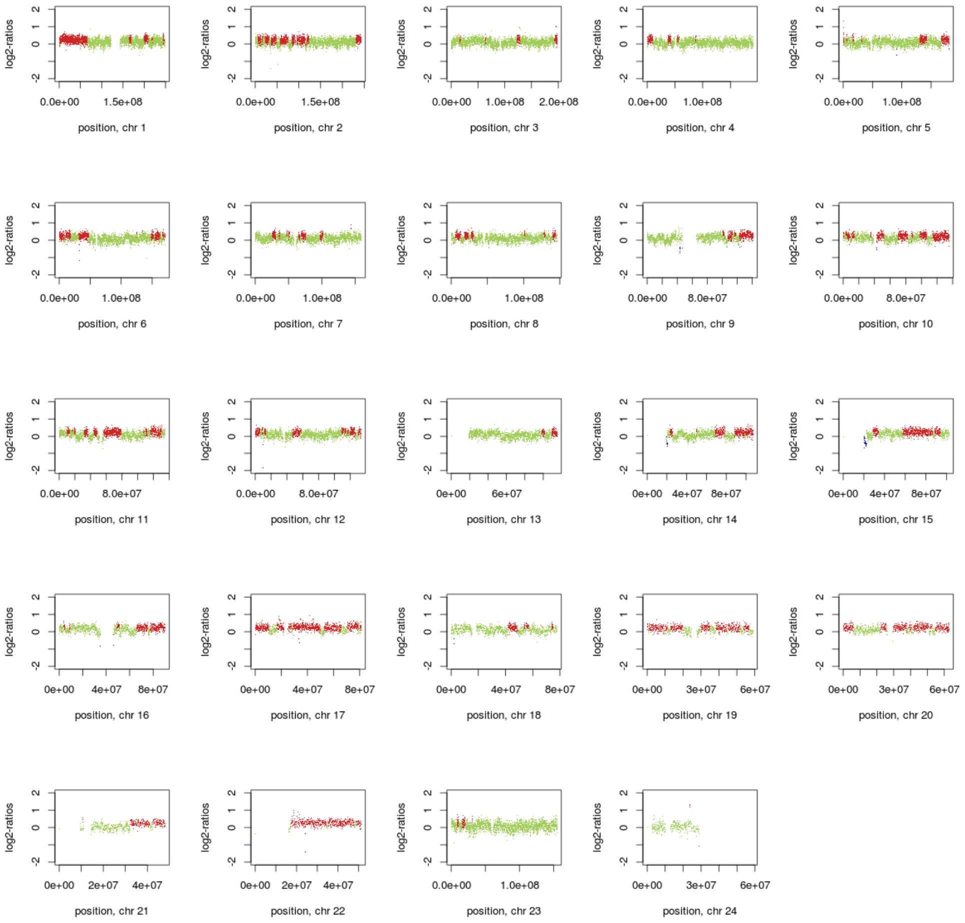


Figure 42: Copy number variation profile for sample *Lunge5*

The regions of all samples with z-scores higher than +3 and lower than -3 are listed in the appendix. Regions which are changed in more than 3 samples were investigated more detailed. At first the start and end point of each changed region were entered in the UCSC Genome Browser and so the affected chromosome band was identified. In addition the UCSC Genome Browser listed the genes located in this changed region. Afterwards the COSMIC databank, which contains the most frequently CNV for lung cancer, was used for comparison. In table 18 the common regions of the samples and also the genes, obtained by using the UCSC Genome Browser, which are also mentioned in the top CNV list of the COSMIC databank are listed. Therefore a differentiation between gain and loss of a region was done. In the COSMIC CNV list, the CNVs are listed according to their frequency of occurrence and the number in the bracket in table 18 represents the list position of the respective CNV. As shown in table 18 no top 400 CNV was found by comparison.

Table 18: Common regions of samples compared with COSMIC *top CNV* list for lung cancer

Common bands	Common region	Samples	Common CNV (position in COSMIC *top CNV* list for lung cancer)
chr1 (p11.2-q21.1)	121424191bp-142613953bp	Lunge2 (loss), Lunge3 (l), Lunge3.2 (l)	Centromere- region, no genes included
chr1 (p36.13)	16902184bp - 17269135bp	Lunge2 (l), Lunge3.4 (gain), Lunge4 (g)	Gain: CROCC(17356) Loss: CROCC(10199)
chr3 (q26.1)	162627563bp-167103184bp	Lunge3.2 (l) Lunge4 (l), Lunge4.2 (l)	SI (19713), SLITRK3 (19702), BCHE (19410), ZBBX (19680)
chr3 (p24.3-p23)	20259538bp-31699338bp	Lunge2 (l), Lunge3 (l), Lunge3.2 (l), Lunge4.2 (l) (26033777bp-26146634bp)	ZNF385D(1712), UBE2E2 (1784), UBE2E1 (1837), NKIRAS1(1864), RPL15(1853), NR1D2(1799), THRB(1867), RARB(1829), TOP2B (1858), NGLY1(1849), OXSM (1917), LRRC3B(1948), NEK10(1898), SLC4A7(1865), EOMES(1883), CMC1(1996), AZI2(2013), ZCWPW2(1983), RBMS3(1981), TGFBR2(1964), GADL1(1991), STT3B(2016)
chr5 (q12.1-q12.3)	62780624bp–64033287bp	Lunge2 (l), Lunge3 (l), Lunge3.2 (l)	HTR1A(3620), RNF180(3651), RGS7BP(3650), SREK1IP1(3514)
chr5 (q13.3)	75794513bp - 75963576bp	Lunge1.1 (l), Lunge2 (l), Lunge3 (l)	IQGAP2(2779), F2RL2(2952)
chr5 (q14.1-q31.1)	79928524bp–130634572bp	Lunge2 (l), Lunge3 (l), Lunge3.2 (l)	DHFR(2565), MSH3(2554), RASGRF2(2703), CKMT2(2787), ZCCHC9(2838), ACOT12(2799), SSBP2(2730), ATG10(2856), RPS23(2902), ATP6AP1L(2855), TMEM167A(2860), XRCC4(2784), VCAN(2804), HAPLN1(2768), EDIL3(2750), COX7C(2650), RASA1(2571), CCNH(2605), TMEM161B(2530), MEF2C(2569), CETN3(2662), MBLAC2(2624), POLR3G(2617), LYSMD3(2629), GPR98(2593), ARRDC3(2757), NR2F1(2616), POU5F2(2570), KIAA0825(2652), FAM172A(2561), ANKRD32(2655), MCTP1(2625), FAM81B(2601), TTC37(2669), GPR150(2640), SPATA9(2644), RHOBTB3(2665), C5orf27(2742), ERAP1(2493), ERAP2(2504), LNPEP(2464), RIOK2(2410), ARSK(2672), GLRX(2745), CAST(2475), RFESD(2657), ELL2(2488), PCSK1(2560), ERAP1(2493), LIX1(2467), RGMB(2449), CHD1(2408), FAM174A(2521), ST8SIA4(2633), PAM(2919), GIN1(3127), PPIP5K2(2996), C5orf30(3086), NUDT12(3016), EFNA5(2847), PJA2(2839), FER(2795), FBXL17(2851), TMEM232(2999), MAN2A1(2966), SLC25A46(3118), WDR36(3099), CAMK4(2948), STARD4(3026), EPB41L4A(2977), SRP19(3023), REEP5 (3003), DCP2(3006), TSSK1B(3136), YTHDC2(3148), TSLP(3092), APC(2918), MCC(2990), KCNN2(3150), TRIM36(3107), PGGT1B(2983), FEM1C(3073), TICAM2(3120), TMED7(3104), CDO1(2984), CCDC112(3013), ATG12(3019), AP3S1(3002), COMMD10(2972), SEMA6A(3087), DTWD2(3015), DMXL1(2959), TNFAIP8(3074), HSD17B4(2969), FAM170A(2852), PRR16(2789), FTMT(2853), SRFBP1(2846), LOX(2909), ZNF474(2886), SNCAIP(2901), SNX2(2884), SNX24(2874), PPIC(2954), PRDM6(3029), CEP120(2933), CSNK1G3(2980), ZNF608(3060), GRAMD3(2978), ALDH7A1(3072), C5orf48(3022), LMNB1(2997), MARCH3(3069), C5orf63(3009), MEGF10(3018), CTXN3(2942), PRRC1(3065), SLC12A2(2970), FBN2(2893), SLC27A6(2940), ISOC1(2926), ADAMTS19(3035), CHSY3(3124), CDC42SE2(3000), HINT1(2930), LYRM7(2920)
chr7 (p22.1)	4782029bp–5628849bp	Lunge1.1 (g), Lunge2 (g), Lunge3.4 (g), Lunge4 (g)	FOXK1(2336), RADIL(2333), PAPOLB(2360), MMD2(2366), RBAK(2367), WIPI2(2412), SLC29A4(2436), TNRC18(2435), FBXL18(2443), ACTB(2441)
chr7 (p14.3)	31935287bp-	Lunge1 (l),	Gain: PDE1C(2392) Loss: PDE1C(21896),

	32503776bp	Lunge2 (g), Lunge3 (l)	
chr7 (q11.23)	72414347bp–76328348bp	Lunge2 (g), Lunge3 (g), Lunge3.2 (g), Lunge3.4 (g) (72414347bp-76202880bp)	POM121(4531), NSUN5P2(4529), LOC541473(4733), STAG3L3(4654), STAG3L1(4633), PMS2P5(4439), NSUN5(4812), TRIM50(4727), FKBP6(4729), BAZ1B(4811), FZD9(4809), TBL2(4813), BCL7B(4815), MLXIPL(4810), VPS37D(4722), DNAJC30(4732), STX1A(4725), WBSCR22(4724), STX1A(4725), ABHD11(4726), CLDN3(4721), CLDN4(4807), WBSCR27(4806), WBSCR28(4808), ELN(4720), LIMK1(4438), EIF4H(4516), LAT2(4518), RFC2(4515), CLIP2(4522), GTF2IRD1(4513), GTF2I(4514), NCF1(4512), GTF2IRD2(4519), STAG3L2(4436), HIP1(4520), CCL26(4523), CCL24(4524), RHBDD2(4643), POR(4646), STYXL1(4645), MDH2(4651), SRRM3(4653), HSPB1(4728), YWHAG(4635), SRCRB4D(4639), ZP3(4636), UPK3B(4730), POMZP3(4652), DTX2(4641)
chr7 (q35)	143239405bp–143571691bp	Lunge1 (g), Lunge1.1 (g), Lunge1.2 (g), Lunge1.3(g), Lunge2 (g)	FAM115C(5001), FAM115A(5293)
chr10 (q11.22)	47384970bp–47743030bp	Lunge2 (g), Lunge3 (g), Lunge3.2 (g), Lunge4 (g) Lunge3.4 (g) (47384970-47675520) Lunge4.2 (g)	ANTXRL(9843)
chr11 (p11.12-p11)	49450722bp–54934099bp	Lunge1.2 (l), Lunge3.2 (l), Lunge4 (l)	OR4C13(18166), OR4A5(21108), OR4C46(21143)
chr12 (p13.31)	7979111bp–8092829bp	Lunge1.1 (g), Lunge1.2 (g), Lunge1.3 (g), Lunge2 (g)	SLC2A14(5437), SLC2A3(5624)
chr12 (p11.21)	31205898bp-32342736bp	Lunge2 (g), Lunge3.2 (g), Lunge3.4 (g)	Centromere- region, no genes included
chr12 (q11-q12)	38058342bp–38512027bp	Lunge1.1 (g), Lunge1.2 (g), Lunge1.3 (g), Lunge2 (g)	Centromere- region, no genes included
chr12 (q13.12)	50521287bp-50915821bp	Lunge2 (g), Lunge3.2 (g), Lunge3.4 (g)	LIMA1(10519), FAM186A(10514), LARP4(10517), DIP2B(10633)
chr12 (q13.13)	53237220bp-54478414bp	Lunge2 (g), Lunge3.2 (g), Lunge3.4 (g)	KRT78(10019), KRT8(10010), KRT18(10018), EIF4B(10005), TENC1(10006),SPRYD3(10009), IGFBP6(10004),SOAT2(10008), CSAD(10366), ZNF740(103639, RARG(10364), ITGB7(10362), ESPL1(10541), MFSD5(10537), AAAS(10540), PFDN5(10538), C12orf10(10542), SP7(10535), SP1(10536), AMHR2(10539), PRR13(10526), PCBP2(10523), MAP3K12(10524), TARBP2(10522), NPFF(10518),ATF7(10521), ATP5G2(10509), CALCOCO1(10511), HOXC13(9937), HOXC12(9932), HOXC11(9936), HOXC10(9939), HOXC9(9938), HOXC8(9934), HOXC6(9940), HOXC5(9942), HOXC4(9941)
chr12 (q24.31)	121638835bp-121131329bp	Lunge2 (g), Lunge3 (g), Lunge4 (g)	MLEC(14347), UNC119B(14351), ACADS(14350), HNF1A(14126), C12orf43(14128), OASL(14127), P2RX7(13856)

chr19 (p13.3)	738015bp–5476714bp	Lunge1.3 (g), Lunge3 (g), Lunge3.2 (g), Lunge4 (g)	PALM(23198), PTBP1(23188), AZU1(23195), PRTN3(23187), ELANE(23191), CFD(23180), MED16(23194), KISS1R(23189), ARID3A(23190), WDR18(23184), GRIN3B(23179), CNN2(23185), ABCA7(23186), HMHA1(23033), POLR2E(23038), GPX4(22825), SBNO2(22620), C19orf26(22834), CIRBP(22830), C19orf24(22836), STK11(22829), ATP5D(22835), MIDN(22838), EFNA2(22833), MUM1(23204), NDUFS7(23499), DAZAP1(23502), C19orf25(23498), ADAMTSL5(23497), MEX3D(23663), GAMT(23508), RPS15(23504), APC2(23505), PCSK4(23506), REEP6(23500), PLK5(23494), UQCR11(23660), TCF3(23664), CSNK1G2(23643), MBD3(23662), ONECUT3(23659), ATP8B3(23647), SCAMP4(23641), REXO1(23654), KLF16(23649), ADAT3(23640), BTBD2(23644), MKNK2(23652), IZUMO4(23813), AP3D1(23812), DOT1L(23815), PLEKHJ1(23820), C19orf35(23827), LINGO3(23828), TIMM13(23821), AP3D1(23812), JSRP1(23826), OAZ1(23825), TMPRSS9(23817), GADD45B(23818), SF3A2(23824), AMH(23823), LSM7(23822) ,LMNB2(23819), GNG7(23949), DIRAS1(23951), SLC39A3(23952), SGTA(23953), THOP1(23832), ZNF554(23831), ZNF555(23833), ZNF556(23839), ZNF57(23960), ZNF77(23978), TLE6(23840), TLE2(22700), AES(22859), GNA11(22858), GNA15(23521), S1PR4(23519), NCLN(23359), CELF5(23212), NFIC(23358), DOHH(23837), C19orf77(23835), FZR1(23958), C19orf71(23969), MFSD12(23961), HMG20B(23962), GIPC3(23971), TBXA2R(23964), PIP5K1C(23836), TJP3(23977), APBA3(23975), MRPL54(23974), RAX2(23973), MATK(23972), ZFR2(23666), ATCAY(23675), DAPK3(23674), EEF2(23677), PIAS4(23670), ZBTB7A(23671), MAP2K2(23673), CREB3L3(23513), ANKRD24(23514), CCDC94(23517), TMIGD2(23518), STAP2(23679), SH3GL1(23842), SIRT6(23509), EBI3(23511), SHD(23515), FSD1(23516), CHAF1A(23520), UBXN6(23523), TNFAIP8L1(23228), MPND(23667), UBXN6(23523), PLIN4(23526), PLIN5(23361), SEMA6B(23363), LRG1(23362), C19orf10(23229), DPP9(23230), FEM1A(23234), TICAM1(23232), UHRF1(23065), ARRDC5(23233), KDM4B(22861), PTPRS(22862), ZNRF4(23372)
chr19 (p11-q11)	24456884bp–28083586bp	Lunge2 (l), Lunge3 (l), Lunge4 (l)	Centromere- region, no genes included
chrX (q23)	110031378bp–110540775bp	Lunge1.2 (l), Lunge2 (l), Lunge3 (l)	CHRDL1(850), PAK3(828), CAPN6(837), DCX(842)
chrX (p11.22- q23)	54350347bp–114905598bp	Lunge2 (l), Lunge3 (l), Lunge4 (l)	WNK3(444), TSR2(449), FGD1(447), GNL3L(469), MAGED2(453), TRO(466), PFKFB1(450) , APEX2(464), ALAS2(451), PAGE2B(460), PAGE2(471), FAM104B(452), PAGE5(457), PAGE3(459), MAGEH1(474), USP51(475), FOXR2(481), RRAGB(480), KLF8(494), UBQLN2(548), UQCRBP1(544), SPIN2B(510), SPIN3(500), FAAH2(476), ZXDB(504), ZXDA (485), SPIN4(484), ARHGEF9(472), ASB12(502), MTMR8(496), ZC4H2(465), ZC3H12B(479), LAS1L(486), VSIG4(509), MSN(498), HEPH(473), EDA2R(546), OPHN1(543), YIPF6(556), STARD8(564), FAM155B (550), AWAT2(534), OTUD6A(535), IGBP1(527), DGAT2L6(530), AWAT1(517), P2RY4(533), ARR3(531), RAB41(523), PDZD11(538), KIF4A(526), GDPD2(519), DLG3(520), TEX11(488), SLC7A3(490), FOXO4(525), CXorf65(539), IL2RG(540), MED12(492), NLGN3(487), ZMYM3(493), NONO(529), ITGB1BP2(515), TAF1(516), INGX(542), OGT(549), ACRC(553), CXCR3(563), NHSL2(572), RPS26P11(589), RGAG4(586), FLJ44635(571), ERCC6L(587), CITED1(599), HDAC8(582), PHKA1(585), DMRTC1B(611), DMRTC1(612), AR(554), EFNB1(567), PJA1(575), EDA(521), SNX12(532), NAP1L6(608), NAP1L2(604), KIAA2022(606), ZDHHC15(620), GJB1(507), RPS4X(601), CHIC1(607), ZCCHC13(605), SLC16A2(602), ABCB7(603), UPRT(628), TTC3P1(619), MAGEE2(622), CDX4(616), RLIM(596), ABCB7(603), MAGEE1(552), FGF16(588), ATRX(598), MAGT1(578), COX7B(576), ATP7A(577), LPAR4(613), PGAM4(592), PGK1 (579), TAF9B(574), CYSLTR1(583), ZCCHC5(614), P2RY10(625), GPR174(637), ITM2A(600), TBX22(641), FAM46D(681), HMGN5(682), BRWD3(644), SH3BGRL (643), POU3F4(676), CYLC1(701), RPS6KA6(714), HDX(715), APOOL(743), SATL1(741), ZNF711 (748), POF1B(750), DACH2(702), CHM(867), KLHL4(798),

			CPXCR1(806), TGIF2LX(3771), PABPC5(3165), PCDH11X(3502), NAP1L3(1110), FAM133A(1023), DIAPH2(687), RPA4(664), PCDH19(646), TNMD(630), TSPAN6(659), SRPX2(657), CSTF2 (660), NOX1(639), XKRX(640), ARL13A(705), TRMT2B(631), TMEM35(651), CENPI(653), DRP2(623), TAF7L(635), TIMM8A(650), BTK(652), RPL36A(649), GLA(638), HNRNPH2(661), ARMCX4(707), ARMCX1(689), ARMCX6(666), ARMCX3(668), ARMCX2(693), NXF5(665), ZMAT1(706), TCEAL2(721), TCEAL6(737), BEX5(734), NXF2(735), NXF2B(695), TMSB15A(673), ARMCX5(663), GPRASP1(686), GPRASP2(703), BHLHB9(696), RAB40AL(704), TCEAL8(728), TCEAL5(732), NGFRAP1(712), RAB40A(692), TCEAL3(709), MORF4L2(697), TMEM31(672), TMSB15B(723), FAM199X(710), IL1RAPL2(627), BEX5 (734), NXF4(711), BEX1(722), NXF3(725), BEX4(716), BEX2(726), TCEAL7(736), WBP5(731), TCEAL4(700), TCEAL1(667), GLRA4(684), PLP1(690), RAB9B(662), H2BFWT(757), H2BFM(774), TEX13A(669), SERPINA7(729), CLDN2(761), NRK(636), MUM1L1(738), CXorf57(758), RNF128(766), TBC1D8B(782), RIPPLY1(769), MORC4(772), RBM41(767), NUP62CL(778), FRMPD3(780), PRPS1(759), TSC22D3(789), MID2(800), TEX13B(788), VSIG1(786), PSMD10(784), ATG4A(792), COL4A6(754), COL4A5(785), NXT2(848), IRS4(831), GUCY2F(858), KCNE1L(809), ACSL4(822), TMEM164(857), AMMECR1(844), RGAG1(810), CHRDL1(850), PAK3(828), CAPN6(837), ALG13(870), TRPC5(887), ZCCHC16(888), DCX(842), LHFPL1(929), AMOT(947), HTR2C(886), LRCH2(1057), LUZP4 (1079), RBMXL3(1080), PLS3(1006), IL13RA2(1037)

6 Discussion

Cancer is a heterogeneous and complex disease and both, the treatment response and the disease susceptibility are affected by the genetic variation [Ziegler *et al.* (2012)]. Especially the heterogeneity of cancer complicates the design of a uniformly effective treatment and hence one major aim of cancer medicine is to move from fixed treatment regime to a therapy tailored to a patient's individual tumor called personalized medicine [Heitzer *et al.* (2013c)].

The identification of biomarkers is a key event for the realization of personalized medicine in cancer therapy [Cima *et al.* (2011); Vogelstein *et al.* (2013)]. In general, such biomarkers can be used for the detection of the disease at an early stage, for the prediction, who will develop cancer, and for the selection of targeted therapies [Cima *et al.* (2011); Ziegler *et al.* (2012)]. The use of tumor tissue biopsies do not enable the detection of biomarker for personalized therapy and only provides a snapshot of mutations at a given time and location [Forshew *et al.* (2012); Heitzer *et al.* (2013c)]. Hence, blood- based assays are of great importance and many studies have shown that there is the possibility for identification of CNV out of cfDNA [Heitzer *et al.* (2013c); Schwarzenbach *et al.* (2011); Leary *et al.* (2012); Heitzer *et al.* (2013b)]. These cfDNA in plasma or serum can for example be used as a noninvasive 'liquid biopsy' to identify prognostic, predictive and pharmacokinetic biomarkers [Dawson *et al.* (2013); Heitzer *et al.* (2013a); Ziegler *et al.* (2012)]. Compared to conventional biopsy, a liquid biopsy has several advantages: It enables a minimally invasive sample acquisition and because of the possibility to collect repeated samples the changes in cfDNA during cancer treatment and during the natural development of the disease can be investigated and assessment of cancer patients after therapy is possible [Heitzer *et al.* (2013c); Schwarzenbach *et al.* (2011); Ghorbian, Ardekani (2012)].

Many studies described a method using massively parallel genomic sequencing of plasma DNA. This kind of sequencing for example is used for the detection of fetal aneuploidy by sequencing plasma DNA derived from the maternal circulation [Heitzer *et al.* (2013c); Chiu, Rossa W. K., Lo, Y. M. Dennis (2013); Chiu, Rossa W. K. *et al.* (2008); Fan *et al.* (2008)] or for detection of chromosomal alterations in the circulation [Chan, K. C. A. *et al.* (2013)]. Such massively parallel genomic sequencing approaches are very time-consuming and prohibitive for routine clinical implementation [Heitzer *et al.* (2013c)].

Heitzer *et al.* (2013c) established a method called 'plasma-Seq' which is based on Illumima's MiSeq instrument to generate whole genome sequencing of plasma DNA to analyze the tumor genomes noninvasively at low cost within 3 days. Therefore shotgun libraries were prepared using the TruSeq DNA LT Sample preparation Kit from Illumina. In general, the procedure happened according to the preparation guide but following modifications were done: i) The fragmentation step was omitted, ii) a lower amount of input DNA was used and iii) a PCR amplification with 25 cycles was done. For this whole-genome sequencing a shallow sequencing depth of about 0.1x was used. The Institute of Human Genetics, Graz received in previous studies useable CNV profiles from this method for patients having breast-, colon- and prostate cancer but till now did not investigate blood samples from lung cancer patients.

Hence, the objective of this thesis was to investigate plasma cfDNA of lung cancer patients, with the aim to identify specific biomarkers, based on the plasma-Seq method. Until now no other study reported the use of the plasma-Seq method for the investigation of plasma samples obtained from patients with lung carcinoma. Hence, only an indirect comparison with results from other types of cancer is possible. The used whole-genome sequencing with a 0.1x coverage is robust and reliable for copy number measurement and a MiSeq run produces a throughput of 3.6 gbp with a read length of 150 bp. Single-read sequencing was performed which involves sequencing DNA from only one end (only the forward strand was sequenced). Advantages of this single-read sequencing are the simple library preparation, the low input DNA requirements, the reducing of the cost and the simplified data analysis [Illumina (2014)]. In addition by using single-read sequencing the sequencing time could be decreased and because of the fact that sequencing of the samples happened untargeted, paired end sequencing was not necessary. As mentioned before advantages of this plasma-Seq method are the reduced costs and the speed of analysis and consequently it is theoretical practicable in clinical implementations. The tumor genomes can be analyzed within 3 days (about 24 hours for library preparation, about 12 hours for sequencing of 150 bp single reads and additional 2.5 hours for bioinformatical analyses) and the analysis of one sample costs about 300 €. But it is to mention, that structural intra- and interchromosomal rearrangements cannot be identified with high confidence by using low coverage whole-genome sequencing. Another short-coming of this plasma-Seq method with such a low coverage is that an identification of mutations of single genes is almost impossible and only copy number variations can be detected. In addition till now it is not possible to say if the changes obtained from the plasma correspond to the primary tumor or to any of the metastatic sites [Heitzer *et al.* (2013c)]. This lack of clarity came about, because at present the exact physiology of the release of tumor DNA is still not well understood and it is unknown whether all tumor cells equally contribute to the plasma DNA. For this plasma-Seq method a percentage of 10% tumor DNA in the blood is enough to achieve good results. This factor is a great advantage of plasma-Seq compared to array CGH.

To reconstruct tumor genomes out of blood from patients with lung cancer, blood was drawn from 5 patients. In addition, a sum of 6 follow up blood samples were taken from these five patients in the course of time. The details on the frequency of sample collection can be seen in table 19 in appendix and was generally low. Also blood from 11 women (F2- F11, F19) and 9 men (M18- M21, M30- M34) without malignant disease were analyzed. For all samples the concentration was determined and quantitative and qualitative analyses as well as whole-genome sequencing using the MiSeq platform (Illumina) were performed. For this study the library preparation was performed according to *TruSeq DNA Sample Preparation Guide (LT)* (Illumina) and *TruSeq Nano DNA Sample Preparation Guide (LT)* (Illumina) with the same modifications used by Heitzer et al (2013). At the beginning of the study three different libraries out of plasma obtained from one and the same patient without a malignant disease were prepared for verification, to test which library preparation protocol was more applicable for this study. The result for sample F19_3 was the most satisfactory result (see chapter 5.2.1) and hence this method was chosen for all other library preparations.

Smoking is the most important risk factor of bronchial carcinoma and the risk increases with quantity and duration of smoking [Cancer Facts and Figures (2012); Bast Jr, Robert C. *et al.* (2000)]. Consistent with this fact four out of five patients having lung cancer are smokers or former smokers, as it can be seen in the patient information in appendix. Because of the early death (five days after first blood collection) it was not possible to find out if patient 5 was smoker, former smoker or non-smoker. Furthermore it can be assumed that lung cancer is no longer more frequently common among men because in this study three out of five patients were female. A further conspicuity is that all patients were in their mid-50s at the first blood collection. This is maybe linked to the fact that people tend to start smoking earlier in their lives.

Many studies described the possibility to use the amount of circulating DNA as a diagnostic value because they showed that the blood of cancer patients possesses a high concentration of cfNA. During tumor progression a release of both, normal (wild-type) DNA and tumor-derived DNA in blood can occur [Schwarzenbach *et al.* (2011)]. In fact also in this study the concentration of cfDNA in the plasma from the cancer patients had an average concentration of 36.4 ng/ml and thus was 9 fold higher than the average concentration of the control samples (4.661 ng/ml) (see table 15 chapter 5.1). However, the amount of circulating DNA as a diagnostic value has been called into question because the plasma cfDNA level can also be increased due to trauma, premalignant states, inflammation, in patients suffering from illnesses and after exercise [Heitzer *et al.* (2013b); Ghorbian, Ardekani (2012)].

As discussed before, till now the exact physiology and rate of release of cfDNA into blood is still not well understood but in literature there are different suggestions for the reason of this event like apoptosis, necrosis and secretion [Schwarzenbach *et al.* (2011); Ghorbian, Ardekani (2012)]. According to Heitzer *et al.* (2013b) the length of DNA released from apoptotic cells is within a size range of 85-230 bp whereas DNA fragments from necrotic cells have sizes larger than 10.000 bp

[Heitzer *et al.* (2013b)]. The plasma samples analyzed in this study had an enrichment of DNA fragments with an average length of about 170 bp and hence, the most of the investigated DNA was probably released from apoptotic cells. The reason for a length of about 170 bp can be described as following: The DNA has to be packed to fit in a cell and therefore the eukaryotic DNA for example is wrapped around special proteins called histones [Nature education (2014)]. A nucleosome consists of eight of these histones and DNA with an average length of 140 bp to 200 bp. The average length of the DNA fragments of about 170 bp measured in this study correspond to the DNA strand length wrapped around a nucleosome plus linker DNA [Breitbach *et al.* (2014)].

Gains or losses of specific chromosomal regions are a hallmark of many cancers. Historically they have been used to identify oncogenes and tumor suppressor targeted by the alterations. Such chromosomal imbalances could be useful as markers of tumorigenesis [Leary *et al.* (2012)]. For this reason segmental z-scores were determined and regions with z-scores higher than +3 and lower than -3 were listed (see appendix). According to literature a z-score under 3 represents the loss of parts of the chromosome or the whole chromosome whereas a z-score higher than 3 represents gains of parts of the chromosome or the whole chromosome. As the table in appendix shows not all samples from one patient (for example Lunge1, Lunge1.1, Lunge1.2, Lunge1.3) showed changes in the same regions. Measuring inaccuracy and the method itself are more probable reasons for this fact than genetic changes. Furthermore, tough the z-score is sensitive; it is not 100% specific. The genome of patient *Lunge1* showed with high probability a germinal mutation on chromosome 7(q35) because all the samples of this patients had a z-score higher than 3 and hence a gain in the region between 143239405 bp to 143571691 bp. In general a germinal mutation occurs in the germ line and hence, to confirm this assumption the genomic DNA obtained from buccal swab has to be analyzed in future.

The highest segmental z-score measured for the lung cancer samples was 28.74 and the segmental z-scores for the control samples except sample *M21* were most of the time under 10 with some outliers up to 20.21. The control sample *M21* showed z-scores up to 153.26 but it is to mention that practically all regions with such a high z-scores are centromere or telomere regions according to USCS Genome Browser. Although normalization of the sequencing data to avoid position effects as wells as variations at centromeres and telomeres was done, artifacts caused by the difficult alignment of reads because of the repetitive sequences in these regions appeared.

To find regions in the genome which are typically changed in lung cancer, regions changed in more than 3 samples were investigated more detailed. In general according to COSMIC database *TP53*, *EGFR* and *KRAS* are the most common genes represented in lung cancer. As mentioned in chapter 5.3 the chromosome band and the genes located in this region were identified by using the UCSC

Genome Browser. Afterwards the COSMIC databank, which contains the most frequent CNV for lung cancer, was used for comparison. As shown in table 18 no top 400 CNV was found by comparison but a detection of hotspots defined by COSMIC would maybe be possible by increasing sample numbers.

In addition to the z-score analysis copy number variation profiles based on the $\log_2$ ratio were generated to illustrate gains and losses of parts of the chromosome or whole chromosome. In general, a positive $\log_2$ ratio represents an amplification (marked with red) whereas a negative $\log_2$ ratio represents a deletion (marked with blue). The cutoff in diagnostics normally is ±0.2. As discussed before it is to consider that because of the unknown ratio of normal DNA and tumor DNA genomic changes may not be detected by a cutoff of ±0.2. As seen in figure 41 the copy number variation profile for sample *Lunge2* was the only one that showed gains and losses of parts of chromosomes. For this sample losses in the chromosome arms 8p and 20p were detected and also a visible, but not marked, decrease of the base line appeared in chromosome arm 4q. In addition, in chromosome arms 1q and 7q gains were detected and a visible, but not marked, increase of the base line appeared in chromosome arm 5p. Although changes were identified, for the fact that the patient died after 5 days since first blood drawing, less changes were detected compared to samples obtained from patients with advanced breast or colon cancer in previous studies of the Institute of Human Genetics, Graz.

All other copy number profiles, excepted *Lunge5* were practically balanced (see appendix). The CNV profile for sample *Lunge5* showed an enormous high number of gains and based on experience it can be said that artifacts during sequencing are quite likely the reason for such a high number of gains. Hence the sequencing results (z-scores and so on) of this sample were not included in the evaluation because of the risk of distorting the results.

There could be different reasons why most of the genomes detected from plasma DNA were balanced and why the genome of patient *Lunge2* showed less changes compared to other cancers but it is to mention that because of the small sample number in this study it is not possible to name the main causes with confidence. Possible reasons could be: i) there exists different events that clear cfDNA from the circulation having a variable half-life in the circulation between 15 minutes and several hours. In this study the first process step (extraction of plasma) was done after 30 min to 2 hours since the blood was drawn and hence, maybe a degradation of DNA happened during this time. In addition Heitzer *et al.* (2013b) determined in their work that not all patients with metastatic disease release a measurable quantity of tumor DNA into the circulation. They mentioned as a possible explanation that maybe because of the short half-life of plasma DNA no significant amount of tumor DNA had been released into the circulation before blood drawing [Heitzer *et al.* (2013b)]. ii) In general, a further - till now not investigated.- reason could be that less tumor DNA is released into blood by a lung tumor compared to other kinds of tumor. iii) Another possible explanation is that less CNVs are present in a lung cancer than in other cancers. In the Progenetix database an overview of copy number abnormalities in human cancer identified by array and chromosomal Comparative Genomic Hybridization (CGH) experiments is given. As seen in figure 43, generated in March 2014, the hotspots of lung cancer were present at most in 30% of all 699 samples. In contrast the gains of chromosome arms 1q and 8q in breast cancer samples were detected in approximately 50% of all 2269 samples (figure 44). At this particular time the assumption that less CNV are present in lung cancer than in other cancers could not be confuted. But the CNV profile generated by Progenetix published in April 2014 including 1732 samples (figure 45) with highly represented CNV called this assumption into question.

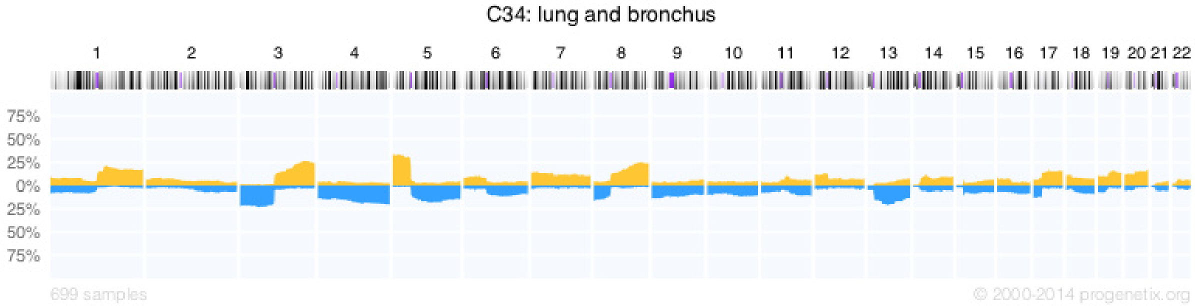


Figure 43: CNV of lung and bronchus carcinoma determined by Progenetix March 2014 [Progenetix (2014)]

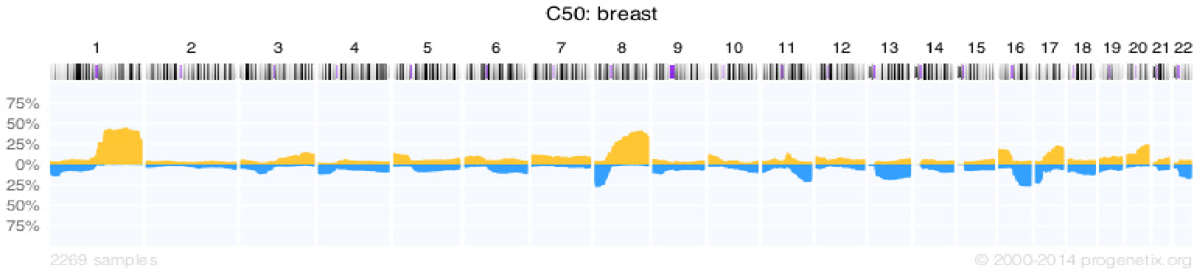


Figure 44: CNV of breast cancer determined by Progenetix [Progenetix (2014)]

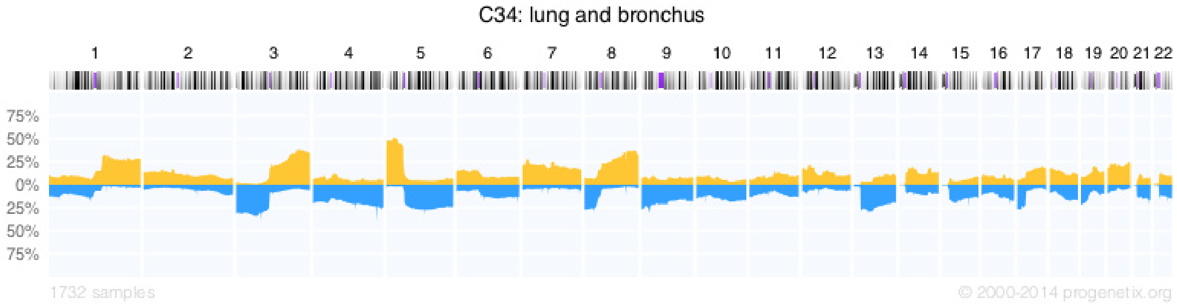


Figure 45: CNV of lung and bronchus carcinoma determined by Progenetix April 2014 [Progenetix (2014)]

By comparison of the results of this study with the results summarized in Progenetix there exist several similarities. As mentioned before in sample *Lunge2* a gain in chromosome arm 1q and a loss in chromosome arm 8p were detected. In addition a visible but not marked decrease of the base line appeared in chromosome arm 4q and a visible, but nor marked, increase of the base line appeared in chromosome arm 5p. All these changes were also detected in a high level by the analyses of the Progenetix database.

In summary, the technology used in this study provides the opportunity to analyze tumor genomes from peripheral blood in detail at different points in time during a disease course as it was shown for sample *Lunge2* and in other studies for breast, colon and prostate cancer performed at the Institute of Human Genetics, Graz. However, due to the fact that this method provided only one copy number profile with detected gains and losses (for sample *Lunge2*) and all other analyzed genomes were balanced, more samples have to be analyzed in the future to draw a conclusion and some modification of the technology has to be done for routine clinical implementation. Modifications such as a direct plasma extraction - immediately after blood drawing at the clinic - to avoid the possibility of degradation of the DNA. With respect to the data analysis, further alterations could be taken into account: Maybe the telomere and centromere regions should be excluded from the analysis to prevent artifacts in these regions and the threshold of ±3 of the z-score should be revised. A gray-value area for example between a z-score of 3 and 10 would be a possibility and only values higher than this threshold should be included in the analyses. But as mentioned before more samples have to be analyzed in future to be able to draw more precise conclusions.

Publication bibliography

Cancer Facts and Figures (2012). In *American Cancer Society* (500812).

Agilent Technologies (2009): Agilent High Sensitivity DNA Kit Quick Start Guide.

Agilent Technologies (2013): Agilent DNA 7500 and DNA 12000 Kit Quick Start Guide.

American Cancer Society (2014): What is non-small cell lung cancer? Available online at http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer, updated on 2/10/2014, checked on 1/14/2014.

Baslan, Timour; Kendall, Jude; Rodgers, Linda; Cox, Hilary; Riggs, Mike; Stepansky, Asya et al. (2012): Genome-wide copy number analysis of single cells. In *Nat Protoc* 7 (6), pp. 1024–1041. DOI: 10.1038/nprot.2012.039.

Bast Jr, Robert C.; Kufe, Donald W.; Pollock, Raphael E.; Weichselbaum, Ralph R.; Holland, James F.; Frei, Emil (2000): Holland-Frei Cancer Medicine. 5th edition. Hamilton (ON): BC Decker.

Bozic, I.; Antal, T.; Ohtsuki, H.; Carter, H.; Kim, D.; Chen, S. et al. (2010): Accumulation of driver and passenger mutations during tumor progression. In *Proceedings of the National Academy of Sciences* 107 (43), pp. 18545–18550. DOI: 10.1073/pnas.1010978107.

Breitbach, Sarah; Tug, Suzan; Helmig, Susanne; Zahn, Daniela; Kubiak, Thomas; Michal, Matthias et al. (2014): Direct Quantification of Cell-Free, Circulating DNA from Unpurified Plasma. In *PLoS ONE* 9 (3), pp. e87838. DOI: 10.1371/journal.pone.0087838.

Brown, T. A. (2002): Genomes. 2nd edition. Oxford: Wiley-Liss.

Brown, T. A. (2010): Gene Cloning & DNA Analysis. An Introduction. 6th edition: Wiley- Blackwell.

Chan, K. C. A.; Jiang, P.; Zheng, Y. W. L.; Liao, G. J. W.; Sun, H.; Wong, J. et al. (2013): Cancer Genome Scanning in Plasma: Detection of Tumor-Associated Copy Number Aberrations, Single-Nucleotide Variants, and Tumoral Heterogeneity by Massively Parallel Sequencing. In *Clinical Chemistry* 59 (1), pp. 211–224. DOI: 10.1373/clinchem.2012.196014.

Charité (2014): Zytogenetische Diagnostik. Edited by Institut für Medizinische Genetik und Humangenetik Berlin. Available online at http://genetik.charite.de/diagnostik/zytogenetik/zytogenetische_diagnostik/, checked on 3/25/2014.

Chiu, Rossa W. K.; Chan, Allen K. C.; Gao, Yuan; Lau, Virginia Y. M.; Zheng, Wenli; Leung, Tak Y. et al. (2008): Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. In *Proceedings of the National Academy of Sciences* 105 (51), pp. 20458–20463. DOI: 10.1073/pnas.0810641105.

Chiu, Rossa W. K.; Lo, Y. M. Dennis (2013): Clinical applications of maternal plasma fetal DNA analysis: translating the fruits of 15 years of research. In *Clinical Chemistry and Laboratory Medicine* 51 (1). DOI: 10.1515/cclm-2012-0601.

Chow, AmyY. (2010): Cell Cycle Control by Oncogenes and Tumor Suppressors: Driving the Transformation of Normal Cells into Cancerous Cells. In *Nature Education*. Available online at http://www.nature.com/scitable/topicpage/cell-cycle-control-by-oncogenes-and-tumor-14191459#, checked on 1/21/2014.

Cima, I.; Schiess, R.; Wild, P.; Kaelin, M.; Schuffler, P.; Lange, V. et al. (2011): Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. In *Proceedings of the National Academy of Sciences* 108 (8), pp. 3342–3347. DOI: 10.1073/pnas.1013699108.

Clancy, Suzanne (2008): Copy number variations (CNVs) have been linked to dozens of human diseases, but can they also represent the genetic variation that was so essential to our evolution? In *Nature Education* 1(1):95. Available online at http://www.nature.com/scitable/topicpage/copy-number-variation-445, checked on 2/5/2014.

Cooper, G. M. (2000): The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates.

Dawson, Sarah-Jane; Tsui, Dana W.Y.; Murtaza, Muhammed; Biggs, Heather; Rueda, Oscar M.; Chin, Suet-Feung et al. (2013): Analysis of Circulating Tumor DNA to Monitor Metastatic Breast Cancer. In *N Engl J Med* 368 (13), pp. 1199–1209. DOI: 10.1056/NEJMoa1213261.

derStandard (2013): Lungenkrebs: Neue Klassifikation soll Therapie verbessern. Available online at http://derstandard.at/1381370516385/Lungenkrebs-Neue-Klassifikation-soll-Therapie-verbessern, updated on 10/31/2013, checked on 4/27/2014.

Dorak, Tevfik M. (2006): Real-time PCR: Taylor & Francis Group (BIOS Advanced Methods).

EPICENTRE: Protocol for MasterAmp™ Buccal Swab DNA Extraction Solution.

Fan, Christina H.; Blumenfeld, Yair J.; Chitkara, Usha; Hudgins, Louanne; Quake, Stephen R. (2008): Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. In *Proceedings of the National Academy of Sciences* 105 (42). DOI: 10.1073/pnas.0808319105.

Fiegler, H.; Geigl, J. B.; Langer, S.; Rigler, D.; Porter, K.; Unger, K. et al. (2007): High resolution array-CGH analysis of single cells. In *Nucleic Acids Research* 35 (3), pp. e15. DOI: 10.1093/nar/gkl1030.

Forshew, Tim; Murtaza, Muhammed; Parkinson, Christine; Gale, Davina; Tsui, Dana W. Y.; Kaper, Fiona et al. (2012): Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA. In *Science Translational Medicine* 4 (136). DOI: 10.1126/scitranslmed.3003726.

Ghorbian, Saeid; Ardekani, Ali M. (2012): Non-Invasive Detection of Esophageal Cancer using Genetic Changes in Circulating Cell-Free DNA. In *Avicenna Journal of Medical Biotechnology* 4 (1), pp. 3–13.

Gibb, Ewan A.; Brown, Carolyn J.; Lam, Wan L. (2011): The functional role of long non-coding RNA in human carcinomas. In *Mol Cancer* 10 (1), p. 38. DOI: 10.1186/1476-4598-10-38.

Hanahan, Douglas; Weinberg, Robert A. (2011): Hallmarks of Cancer: The Next Generation. In *Cell* 144 (5), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013.

Heitzer, E.; Auer, M.; Gasch, C.; Pichler, M.; Ulz, P.; Hoffmann, E. M. et al. (2013a): Complex Tumor Genomes Inferred from Single Circulating Tumor Cells by Array-CGH and Next-Generation Sequencing. In *Cancer Research* 73 (10), pp. 2965–2975. DOI: 10.1158/0008-5472.CAN-12-4140.

Heitzer, Ellen; Auer, Martina; Hoffmann, Eva Maria; Pichler, Martin; Gasch, Christin; Ulz, Peter et al. (2013b): Establishment of tumor-specific copy number alterations from plasma DNA of patients with cancer. In *Int. J. Cancer* 133 (2), pp. 346–356. DOI: 10.1002/ijc.28030.

Heitzer, Ellen; Ulz, Peter; Belic, Jelena; Gutschi, Stefan; Quehenberger, Franz; Fischereder, Katja et al. (2013c): Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. In *Genome Med* 5 (4), p. 30. DOI: 10.1186/gm434.

Herbst, R. S.; Heymach, J. V.; Lippman, S. M. (2008): Lung cancer. In *N Engl J Med* 359. DOI: 10.1056/NEJMra0802714.

Hupe, P.; Stransky, N.; Thiery, J.-P.; Radvanyi, F.; Barillot, E. (2004): Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. In *Bioinformatics* 20 (18), pp. 3413–3422. DOI: 10.1093/bioinformatics/bth418.

Illumina: TruSeq Nano DNA Sample Prep Guide 15041110 A 2013.

Illumina (2010): Illumina Sequencing Technology. Highest data accuracy, simple workflow, and a broad range of applications. Available online at http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf.

Illumina (2012a): TruSeq DNA Sample Preparation Guide.

Illumina (2012b): MiSeq System User Guide.

Illumina (2013): Preparing Libraries for Sequencing on the MiSeq.

Illumina (2014): Single-Read Sequencing. Available online at http://www.illumina.com/technology/single_read_sequencing_assay.ilmn, checked on 4/2/2014.

Invitrogen (2010a): Qubit 2.0 Fluorometer User Manual (MAN0003231).

Invitrogen (2010b): Qubit™ dsDNA HS Assay Kits (MAN0002326).

Jemal, Ahmedin; Bray, Freddie; Center, Melissa M.; Ferlay, Jacques; Ward, Elizabeth; Forman, David (2011): Global cancer statistics. In *CA: A Cancer Journal for Clinicians* 61 (2), pp. 69–90. DOI: 10.3322/caac.20107.

Kranenburg, Onno (2005): The KRAS oncogene: Past, present, and future. In *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1756 (2), pp. 81–82. DOI: 10.1016/j.bbcan.2005.10.001.

Leary, Rebecca J.; Sausen, Mark; Kinde, Isaac; Papadopoulos, N.; Carpten, John D.; Craig, David et al. (2012): Detection of Chromosomal Alterations in the Circulation of Cancer Patients with Whole-Genome Sequencing. In *Science Translational Medicine* 4 (162), pp. 162ra154. DOI: 10.1126/scitranslmed.3004742.

Loman, Nicholas J.; Misra, Raju V.; Dallman, Timothy J.; Constantinidou, Chrystala; Gharbia, Saheer E.; Wain, John; Pallen, Mark J. (2012): Performance comparison of benchtop high-throughput sequencing platforms. In *Nat Biotech* 30 (5), pp. 434–439. Available online at http://dx.doi.org/10.1038/nbt.2198.

Luo, Susan Y.; Lam, David C. L. (2013): Oncogenic driver mutations in lung cancer. In *Translational Respiratory Medicine. DOI:* 10.1186/2213-0802-1-6.

Lynch, Jennifer R.; Brown, Jennifer M. (1990): The polymerase chain reaction: current and future clinical applications. In *Journal of Medical Genetics* 27, pp. 2–7.

McGranahan, Nicholas; Burrell, Rebecca A.; Endesfelder, David; Novelli, Marco R.; Swanton, Charles (2012): Cancer chromosomal instability: therapeutic and diagnostic challenges. In *EMBO Rep* 13 (6), pp. 528–538. DOI: 10.1038/embor.2012.61.

Mechanic, L. E.; Bowman, E. D.; Welsh, J. A.; Khan, M. A.; Hagiwara, N.; Enewold, L. et al. (2007): Common Genetic Variation in TP53 Is Associated with Lung Cancer Risk and Prognosis in African Americans and Somatic Mutations in Lung Tumors. In *Cancer Epidemiology Biomarkers & Prevention* 16 (2), pp. 214–222. DOI: 10.1158/1055-9965.EPI-06-0790.

Mörike, Klaus D.; Betz, Eberhard; Mergenthaler (1989): Biologie des Menschen. Heidelberg Wiesbaden: Quelle & Meyer Verlag.

Mutschler, Ernst; Schaible, Hans-Georg; Vaupel, Peter (2007): Anatomie Physiologie Pathophysiologie des Menschen: Wissenschaftliche Verlagsgesellschaft mbH Stuttgart.

National Collaborating Centre for Cancer (UK) (2011): The Diagnosis and Treatment of Lung Cancer (Update). Cardiff (UK): National Collaborating Centre for Cancer (UK): NICE Clinical Guidelines (101).

Nature education (2014): DNA Is a Structure That Encodes Biological Information. Available online at http://www.nature.com/scitable/topicpage/dna-is-a-structure-that-encodes-biological-6493050, checked on 11/20/2013.

NCBI (2014): MET met proto-oncogene [Homo sapiens (human)]. Available online at http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=4233, updated on 4/21/2014, checked on 2/25/2014.

Normanno, Nicola; Luca, Antonella de; Bianco, Caterina; Strizzi, Luigi; Mancino, Mario; Maiello, Monica R. et al. (2006): Epidermal growth factor receptor (EGFR) signaling in cancer. In *Gene* 366 (1), pp. 2–16. DOI: 10.1016/j.gene.2005.10.018.

O'Huallachain, M.; Karczewski, K. J.; Weissman, S. M.; Urban, A. E.; Snyder, M. P. (2012): Extensive genetic variation in somatic human tissues. In *Proceedings of the National Academy of Sciences* 109 (44), pp. 18018–18023. DOI: 10.1073/pnas.1213736109.

Orr, Bernardo; Compton, Duane A. (2013): A Double-Edged Sword: How Oncogenes and Tumor Suppressor Genes Can Contribute to Chromosomal Instability. In *Front. Oncol.* 3. DOI: 10.3389/fonc.2013.00164.

Progenetix (2014): Progenetix - genomic copy number aberrations in cancer. Edited by Michael Baudis. Available online at http://www.progenetix.org/cgi-bin/pgHome.cgi, updated on 4/23/2014, checked on 4/19/2014.

QIAGEN (2008): QIAcube User Manual.

QIAGEN (2010): QIAamp DNA Mini and Blood Mini Handbook.

QIAGEN SABioscience (2008): Real-Time PCR for Systems Biology: A Review on Real- Time PCR-Related Technologies and Their Applications in the Post- Genomic Era. In *Pathways Magazine* 8.

Quail, Michael; Smith, Miriam E.; Coupland, Paul; Otto, Thomas D.; Harris, Simon R.; Connor, Thomas R. et al. (2012): A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. In *BMC Genomics* 13 (1), p. 341. DOI: 10.1186/1471-2164-13-341.

Redon, Richard; Ishikawa, Shumpei; Fitch, Karen R.; Feuk, Lars; Perry, George H.; Andrews, T. Daniel et al. (2006): Global variation in copy number in the human genome. In *Nature* 444 (7118), pp. 444–454. DOI: 10.1038/nature05329.

Schwarzenbach, Heidi; Hoon, Dave S. B.; Pantel, Klaus (2011): Cell-free nucleic acids as biomarkers in cancer patients. In *Nat Rev Cancer* 11 (6), pp. 426–437. DOI: 10.1038/nrc3066.

Shendure, Jay; Aiden, Erez Lieberman (2012): The expanding scope of DNA sequencing. In *Nat Biotechnol* 30, pp. 1084–1094. DOI: 10.1038/nbt.2421.

Shendure, Jay; Ji, Hanlee (2008): Next-generation DNA sequencing. In *Nat Biotechnol* 26 (10), pp. 1135–1145. DOI: 10.1038/nbt1486.

Statistik Austria (2013). Available online at http://www.statistik.at/web_de/statistiken/gesundheit/krebserkrankungen/luftroehre_bronchien_lunge/index.html, updated on 10/31/2013, checked on 11/25/2013.

Stiewe, Thorsten (2007): The p53 family in differentiation and tumorigenesis. In *Nature Reviews Cancer* (7), pp. 165–167. DOI: 10.1038/nrc2072.

Strachan, T.; Read, A. P. (1999): Human Molecular Genetics. 2nd edition. New York: Wiley-Liss.

Theisen, Aaron (2008): Microarray-based Comparative Genomic Hybridization (aCGH). In *Nature Education* 1(1):45. Available online at http://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432, checked on 2/22/2014.

Trust Sanger institute (2012a): COSMIC (Catalogue of Somatic Mutations in Cancer). Available online at http://www.sanger.ac.uk/resources/databases/cosmic.html, updated on 11/29/2012, checked on 4/9/2014.

Trust Sanger institute (2012b): Cosmic » Tissue » Overview »Lung. Available online at http://cancer.sanger.ac.uk/cosmic/browse/tissue#sn=lung&ss=all&hn=all&sh=all&in=t&src=tissue, checked on 2/24/2014.

Vogelstein, B.; Papadopoulos, N.; Velculescu, V. E.; Zhou, S.; Diaz, L. A.; Kinzler, K. W. (2013): Cancer Genome Landscapes. In *Science* 339 (6127), pp. 1546–1558. DOI: 10.1126/science.1235122.

Xiang, Dong; Zhang, Bicheng; Doll, Donald; Shen, Kui; Kloecker, Goetz; Freter, Carl (2013): Lung cancer screening: from imaging to biomarker. In *Biomark Res* 1 (1), p. 4. DOI: 10.1186/2050-7771-1-4.

Ziegler, Andreas; Koch, Armin; Krockenberger, Katja; Großhennig, Anika (2012): Personalized medicine using DNA biomarkers: a review. In *Hum Genet* 131 (10), pp. 1627–1638. DOI: 10.1007/s00439-012-1188-9.

Zong, Chenghang; Lu, Sijia; Chapman, Alec R.; Xie, Sunney X. (2012): Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. In *Science* 338 (6114), pp. 1622–1626. DOI: 10.1126/science.1229164.

List of tables

List of figures

List of abbreviations

µl	microliter
ml	mililiter
pg	picogram
ng	nanogram
g	gram
nM	nanomolar
pM	picpmolar
sec	seconds
min	minutes
°C	degree Celcius
RT	room temperature
DNA	deoxyribonucleic acid
gDNA	genomic deoxyribonucleic acid
RNA	ribonucleic acid
cfDNA	cell free DNA
cfNa	cell free nucleic acid
dNTP	deoxynucleoside triphosphate
bp	base pair
kbp	kilo base pairs
Kb	kilobase
Mb	megabase
Chr	chromosome
SNP	single nucleotide polymorphism
STR	short tandem repeat
CNV	copy number variation
CIN	chromosomal instability

CNP	copy number polymorphism
SNV	single nucleotide variation
CGH	comparative genome hybridization
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
FiSH	fluorescence in situ hybridization
NGS	next generation sequencing
SAM	sequence alignment /map
BWA	Burrous Wheeler Alignment
EDTA	ethylenediaminetetraacetic acid
(x) g	gravitational acceleration
rpm	round per minute
HT1	hybridization buffer
NaOH	sodium hydroxide
EtOH	Ethanol
EML4	echinoderm microtubule-associated protein-like 4
ALK	anaplastic lymphoma kinase
pRb	retinoblastoma protein
EGFR	epidermal growth factor receptor
SCLC	small cell lung cancer
NSCLC	non- small cell lung cancer
CFR	case form report

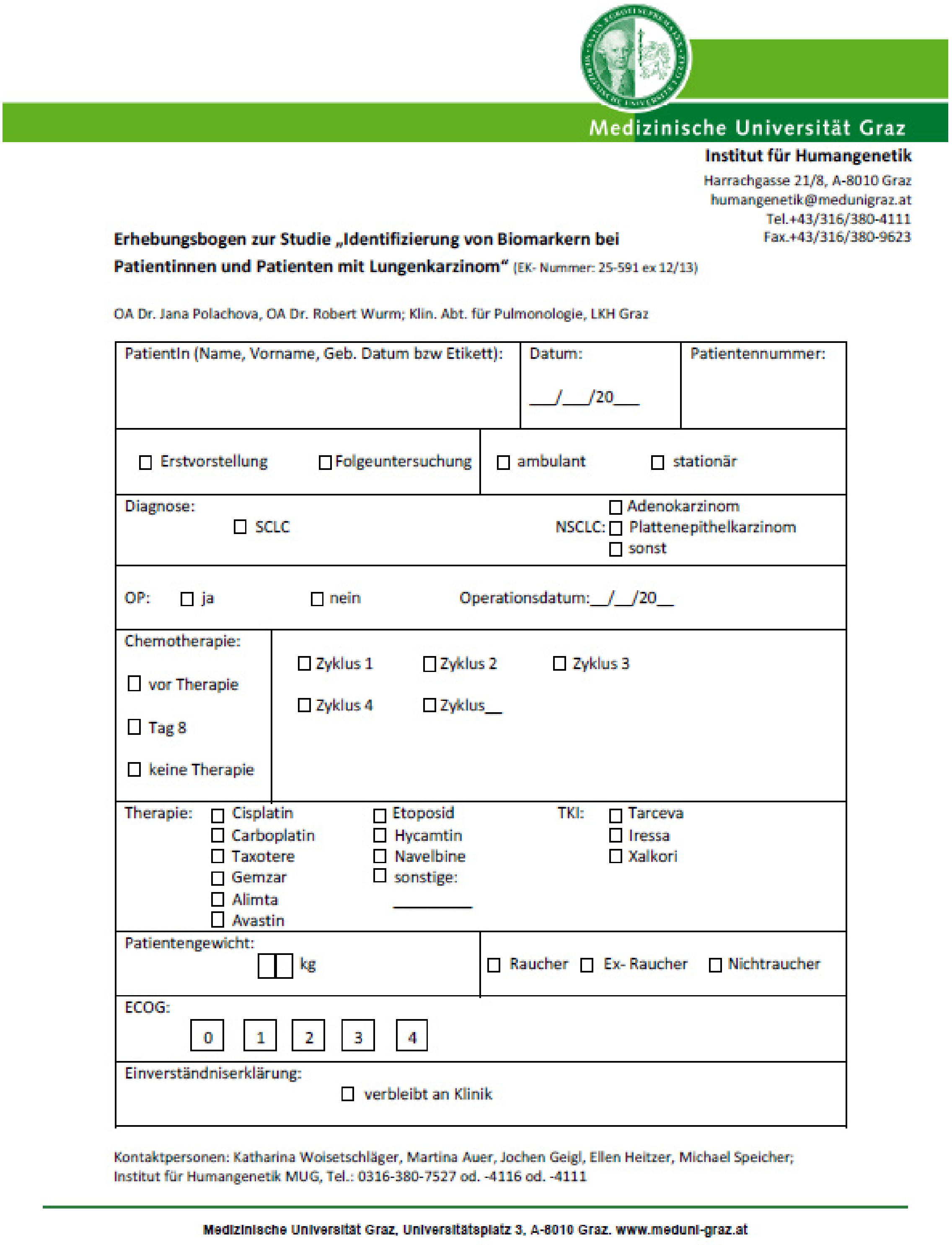
Appendix

Patient Information

Table 19: Patient information

Patienten Name	Sample Name	Date	Sex	Date of Birth	Age	Weight	Histology	Status	Therapie	ECOG	Smoker/ Nonsmoker	Stationary/ambulant	Library Index
B W	Lunge 1	18.11.2013	m	03.12.1959	54	68	NSCLC (Adenocarcinoma)	Before therapie	Cisplatin, Alimta	0	Ex- Smoker	Stationary	5
	Lunge 1.1	28.11.2013				68		Day 10 (Cycle 1)	Cisplatin, Alimta	1		Ambulant	18
	Lunge 1.2	09.12.2013				68		Cycle 2	Cisplatin, Alimta	1		Stationary	19
	Lunge1.3	09.01.2014				68		Cycle 3	Cisplatin, Alimta	1		Ambulant	8
K A	Lunge 2	13.11.2013	f	12.12.1956	57	70	NSCLC (Adenocarcinoma)	Before therapie	/	1			6
	died on 18.11.2013												
G G	Lunge 3	17.12.2013	f	22.11.1957	56	103	SCLC	Before therapie	Cisplatin, Etoposid	0	Smoker	Stationary	1
	Lunge3.2	08.01.2014	f			105		Cycle 2	Cisplatin, Etoposid	0		Stationary	9
	Lunge3.4	28.02.2014				105		Cycle 4	Cisplatin, Etoposid	0			6
W S	Lunge 4	17.12.2013	f	02.10.1958	55	90	NSCLC (Adenocarcinoma)	Before therapie	Cisplatin, Alimta	2	Ex- Smoker	Stationary	3
	Lunge4.2	28.01.2014			55	83		Cycle 2	Carboplatin, Alimta	2		Stationary	12
W H	Lunge5	05.03.2014	m	15.10.1955	58	95	NSCLC (Adenocarcinoma)	Before therapie	Cisplatin, Alimta	0	Ex- Smoker	Stationary	7

Case form report



Medizinische Universität Graz

Institut für Humangenetik

Harrachgasse 21/8, A-8010 Graz
humangenetik@medunigraz.at
Tel.+43/316/380-4111
Fax.+43/316/380-9623

Erhebungsbogen zur Studie „Identifizierung von Biomarkern bei Patientinnen und Patienten mit Lungenkarzinom“ (EK- Nummer: 25-591 ex 12/13)

OA Dr. Jana Polachova, OA Dr. Robert Wurm; Klin. Abt. für Pulmonologie, LKH Graz

PatientIn (Name, Vorname, Geb. Datum bzw Etikett):

Datum: ___/___/20___

Patientennummer:

□ Erstvorstellung □ Folgeuntersuchung □ ambulant □ stationär

Diagnose: □ SCLC NSCLC: □ Adenokarzinom □ Plattenepithelkarzinom □ sonst

OP: □ ja □ nein Operationsdatum:__/__/20__

Chemotherapie:

□ vor Therapie

□ Tag 8

□ keine Therapie

□ Zyklus 1 □ Zyklus 2 □ Zyklus 3

□ Zyklus 4 □ Zyklus__

Therapie:

□ Cisplatin
□ Carboplatin
□ Taxotere
□ Gemzar
□ Alimta
□ Avastin

□ Etoposid
□ Hycamtin
□ Navelbine
□ sonstige: __________

TKI:

□ Tarceva
□ Iressa
□ Xalkori

Patientengewicht: □□ kg

□ Raucher □ Ex- Raucher □ Nichtraucher

ECOG: 0 1 2 3 4

Einverständniserklärung: □ verbleibt an Klinik

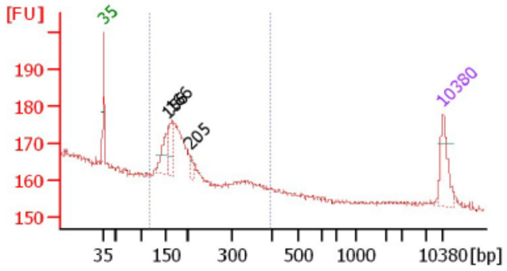
Kontaktpersonen: Katharina Woisetschläger, Martina Auer, Jochen Geigl, Ellen Heitzer, Michael Speicher; Institut für Humangenetik MUG, Tel.: 0316-380-7527 od. -4116 od. -4111

Medizinische Universität Graz, Universitätsplatz 3, A-8010 Graz. www.meduni-graz.at

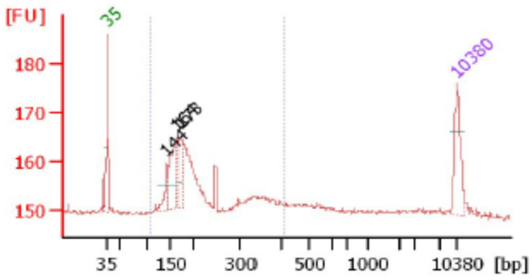
Figure 46: Case form report designed for this study

High sensitivity Bioanalyzer results

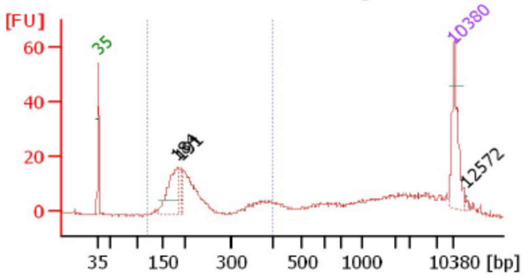
Lunge 1.1



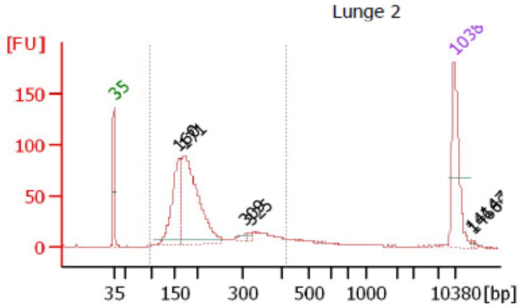
Lunge1.2



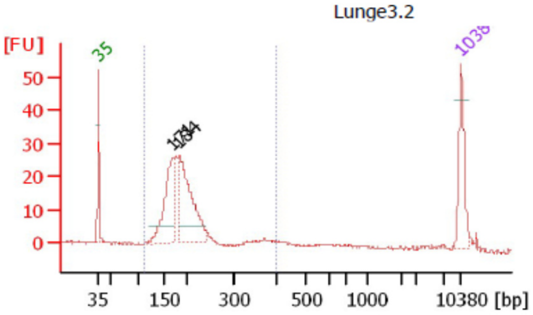
Lunge1.3



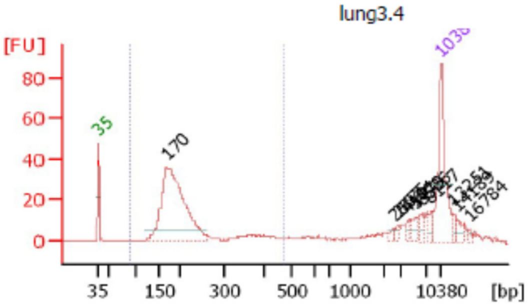
Lunge 2



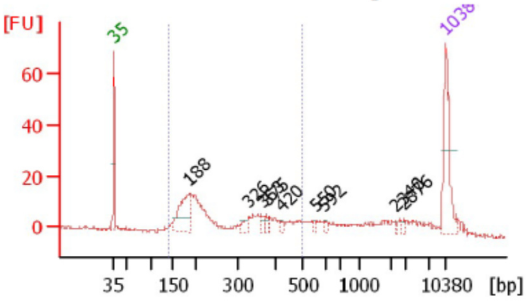
Lunge3.2



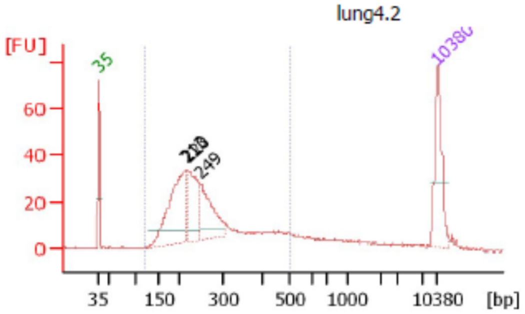
lung3.4



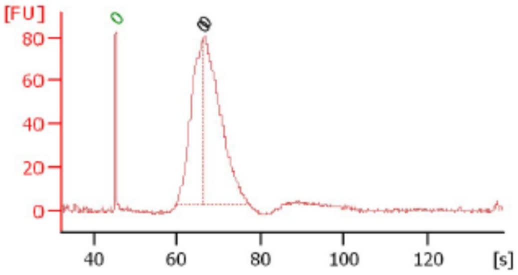
Lunge4



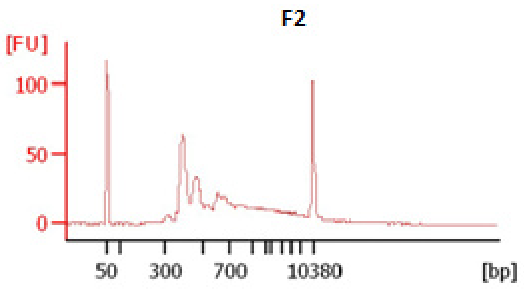
lung4.2

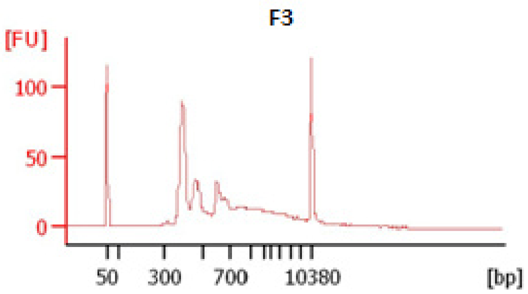


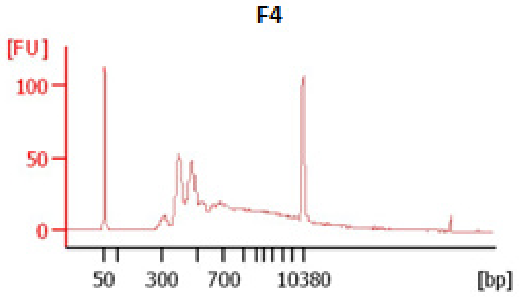
Lung5

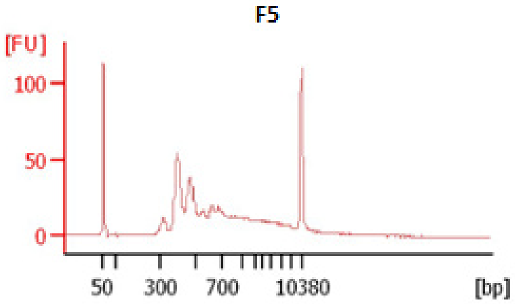


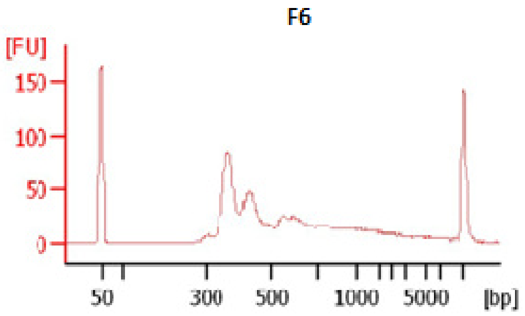
DNA 7500 Bioanalyzer results

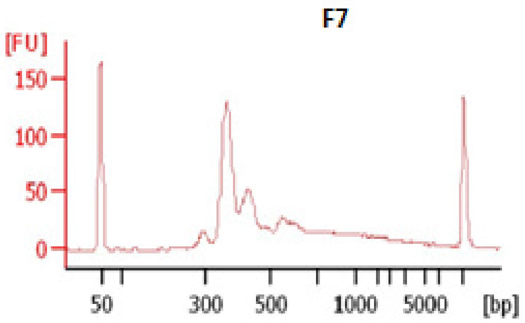

F2

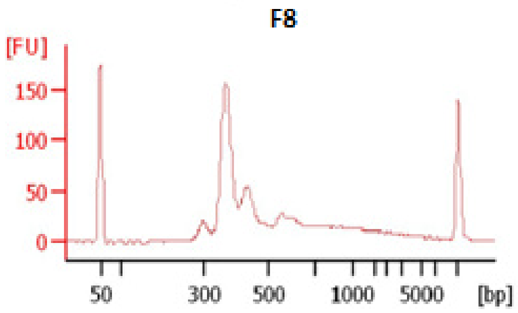

F3

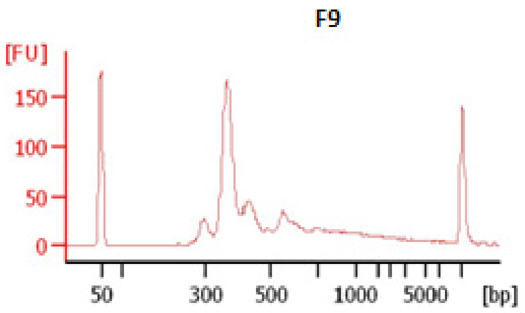

F4


F5

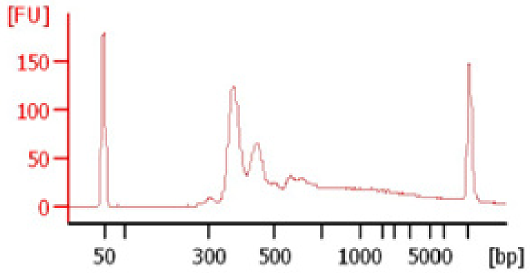

F6


F7

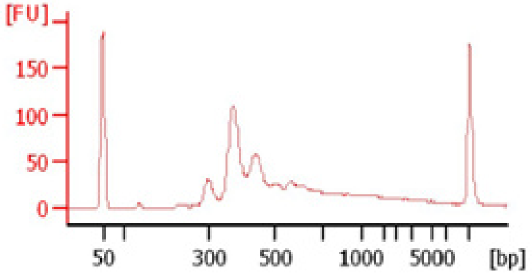

F8


F9

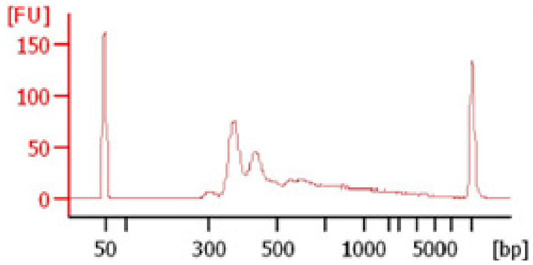
F10



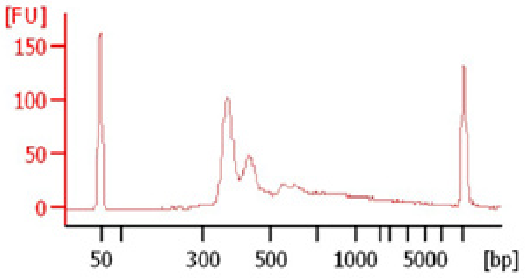
F11



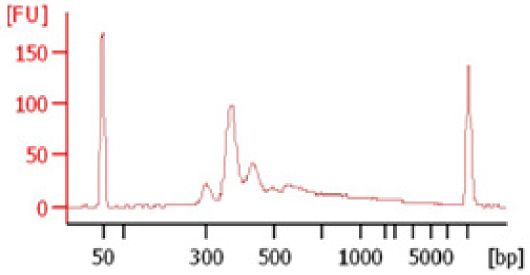
M30



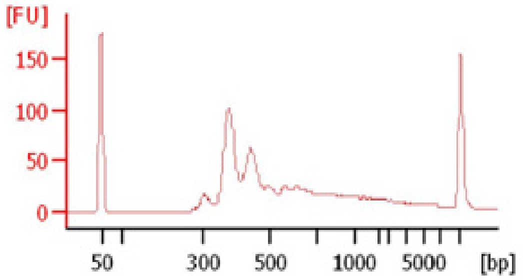
M31



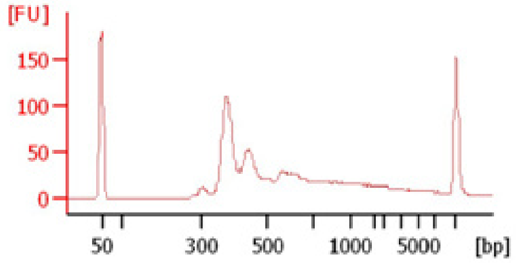
M32



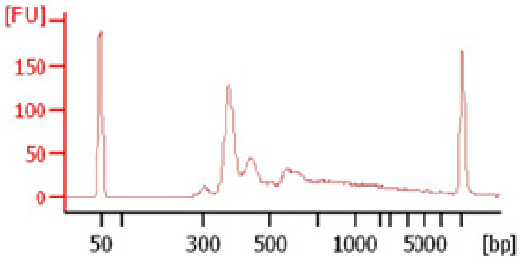
M33



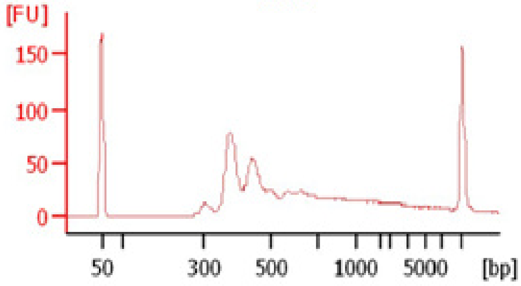
M34



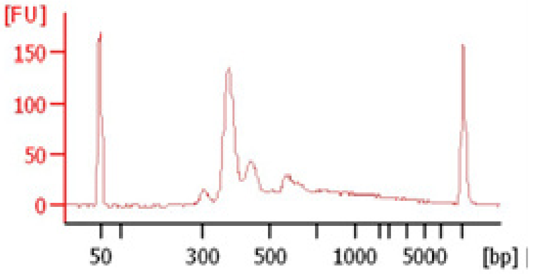
M18



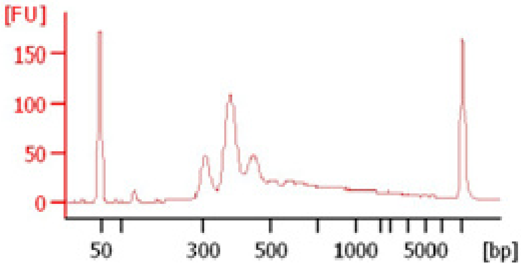
M19



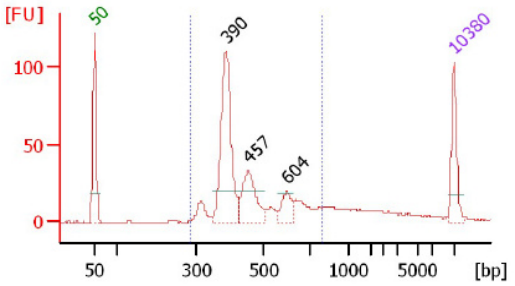
M20



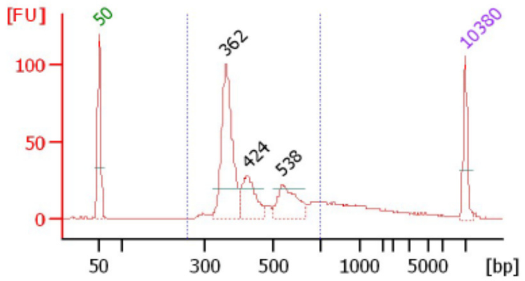
M21



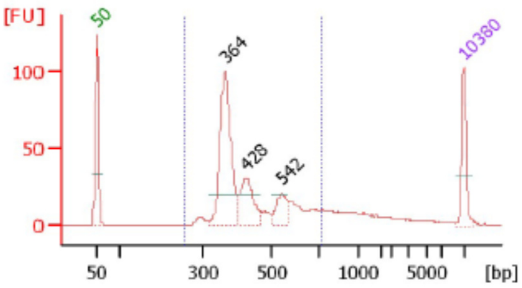
F19



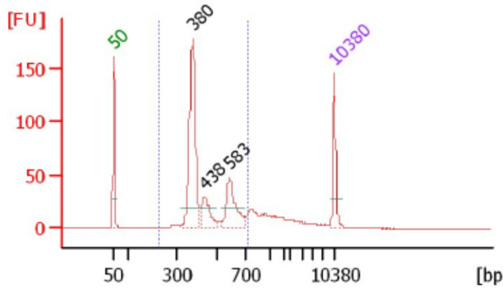
Lunge1.1



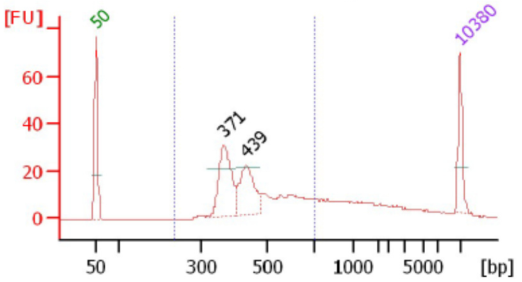
Lunge1.2



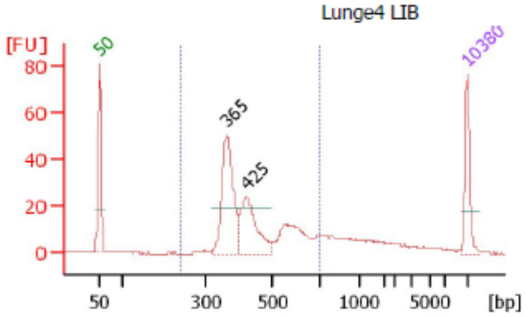
Lunge2LIB



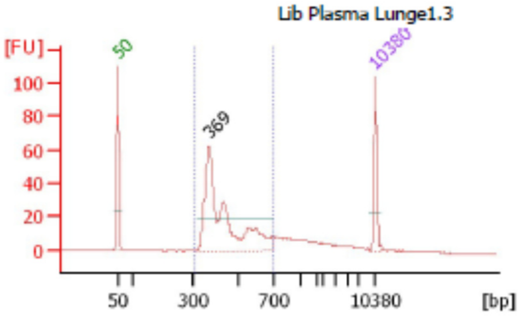
Lunge3 LIB

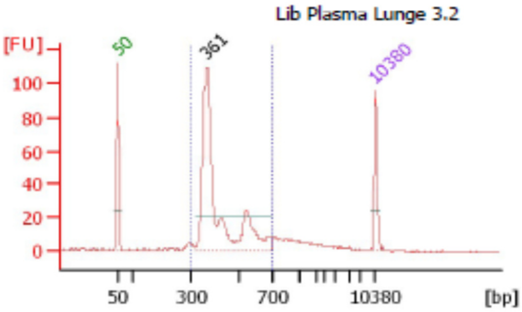


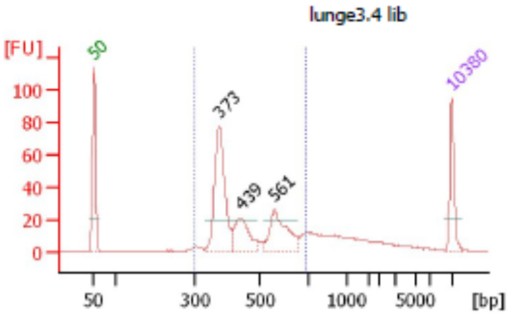
Lunge4 LIB

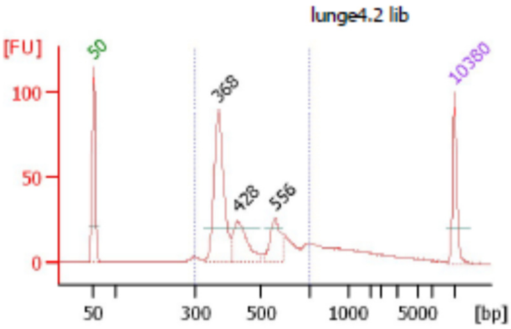


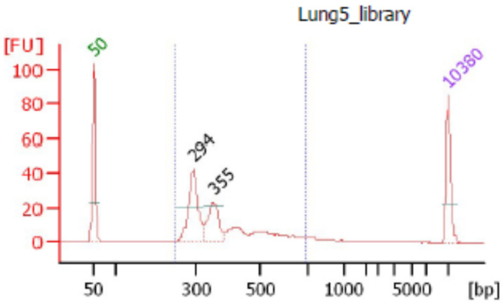
Lib Plasma Lunge1.3




Lib Plasma Lunge 3.2
[FU]
100
80
60
40
20
0
50
361
10380
50
300
700
10380
[bp]


lunge3.4 lib
[FU]
100
80
60
40
20
0
50
373
439
561
10380
50
300
500
1000
5000
[bp]


lunge4.2 lib
[FU]
100
50
0
50
368
428
556
10380
50
300
500
1000
5000
[bp]


Lung5_library
[FU]
100
80
60
40
20
0
50
294
355
10380
50
300
500
1000
5000
[bp]

Sequencing results

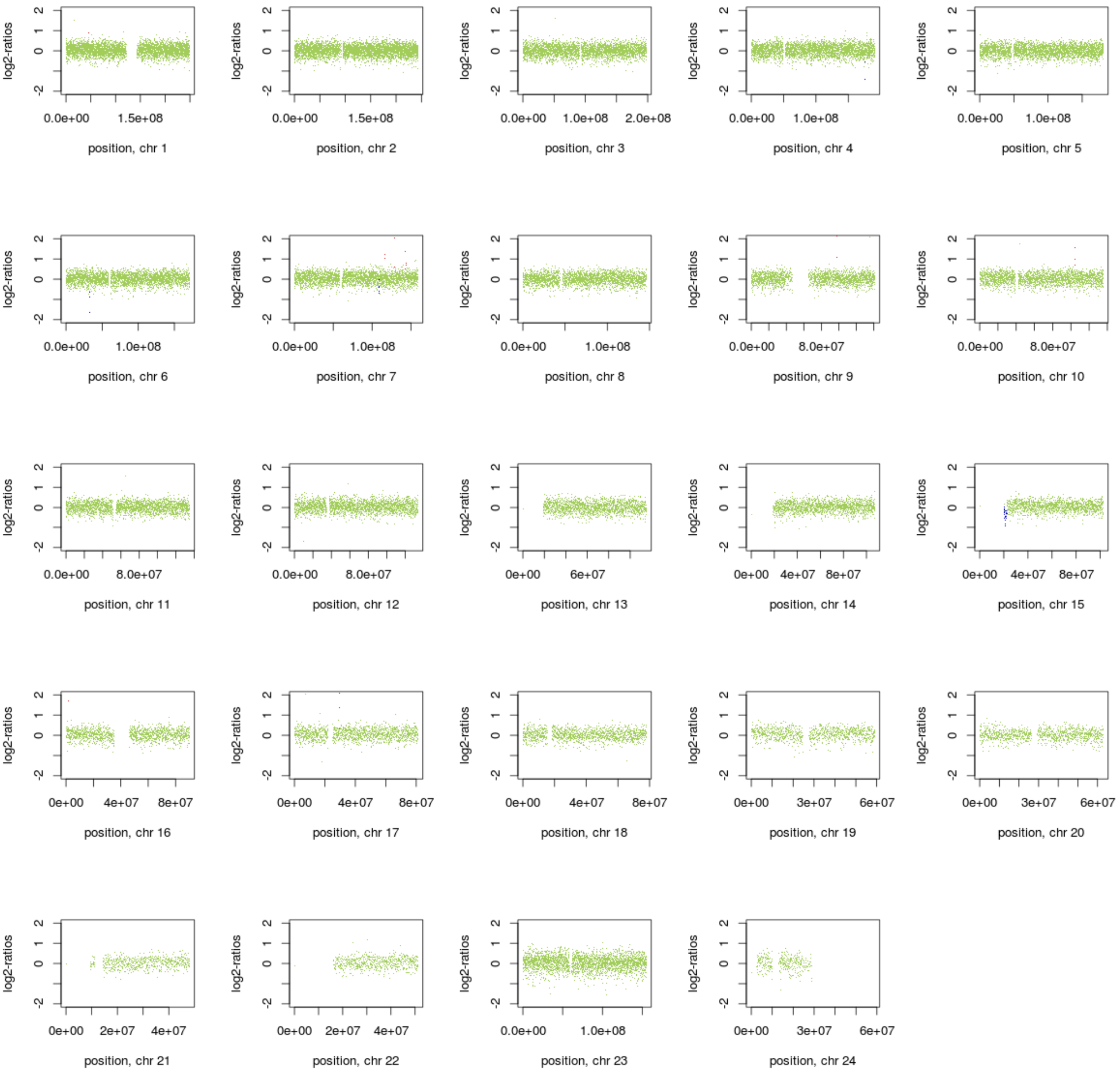


Figure 47: CNV profile for sample *Lunge1*

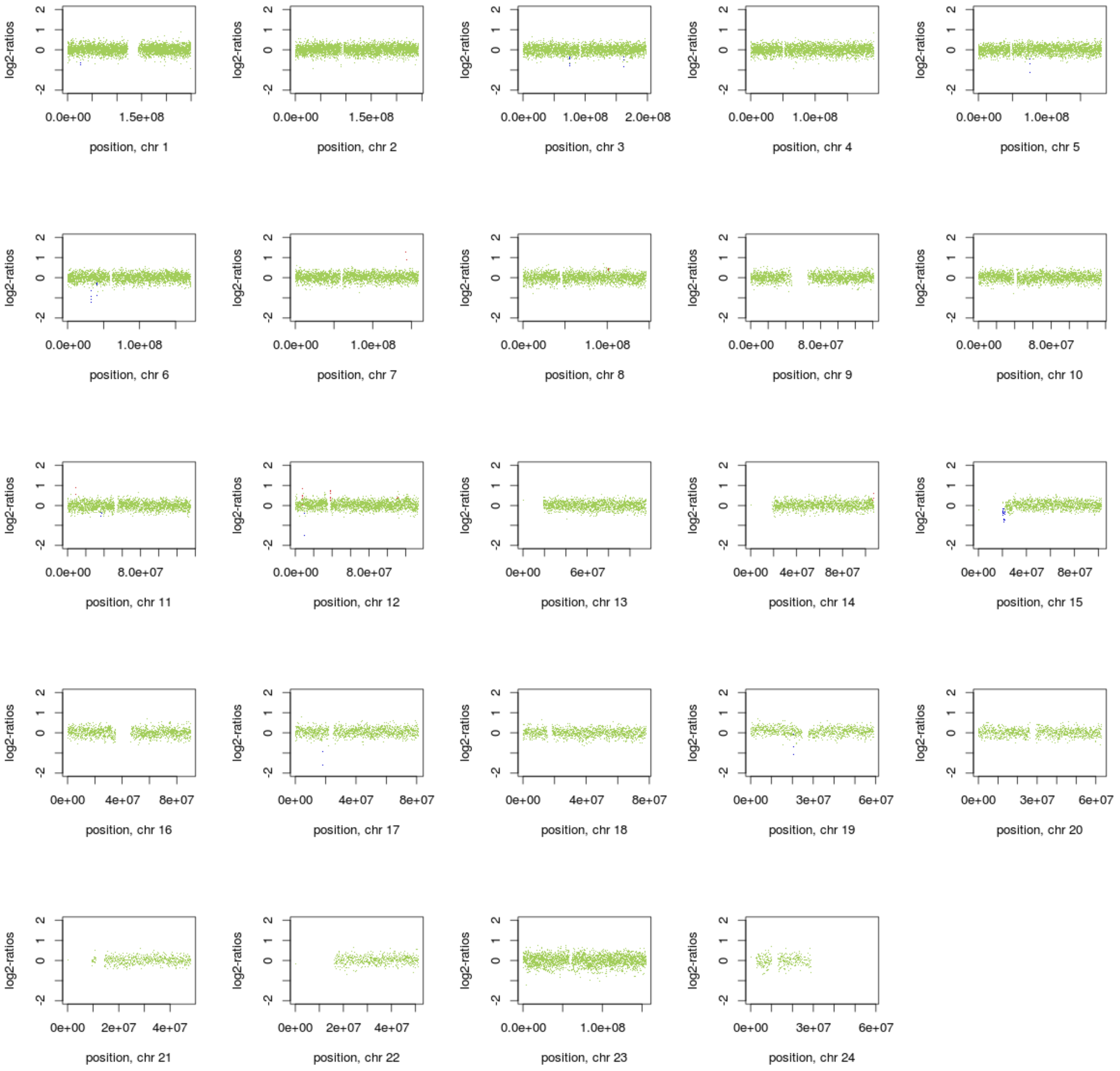


Figure 48: CNV profile for sample *Lunge1.1*

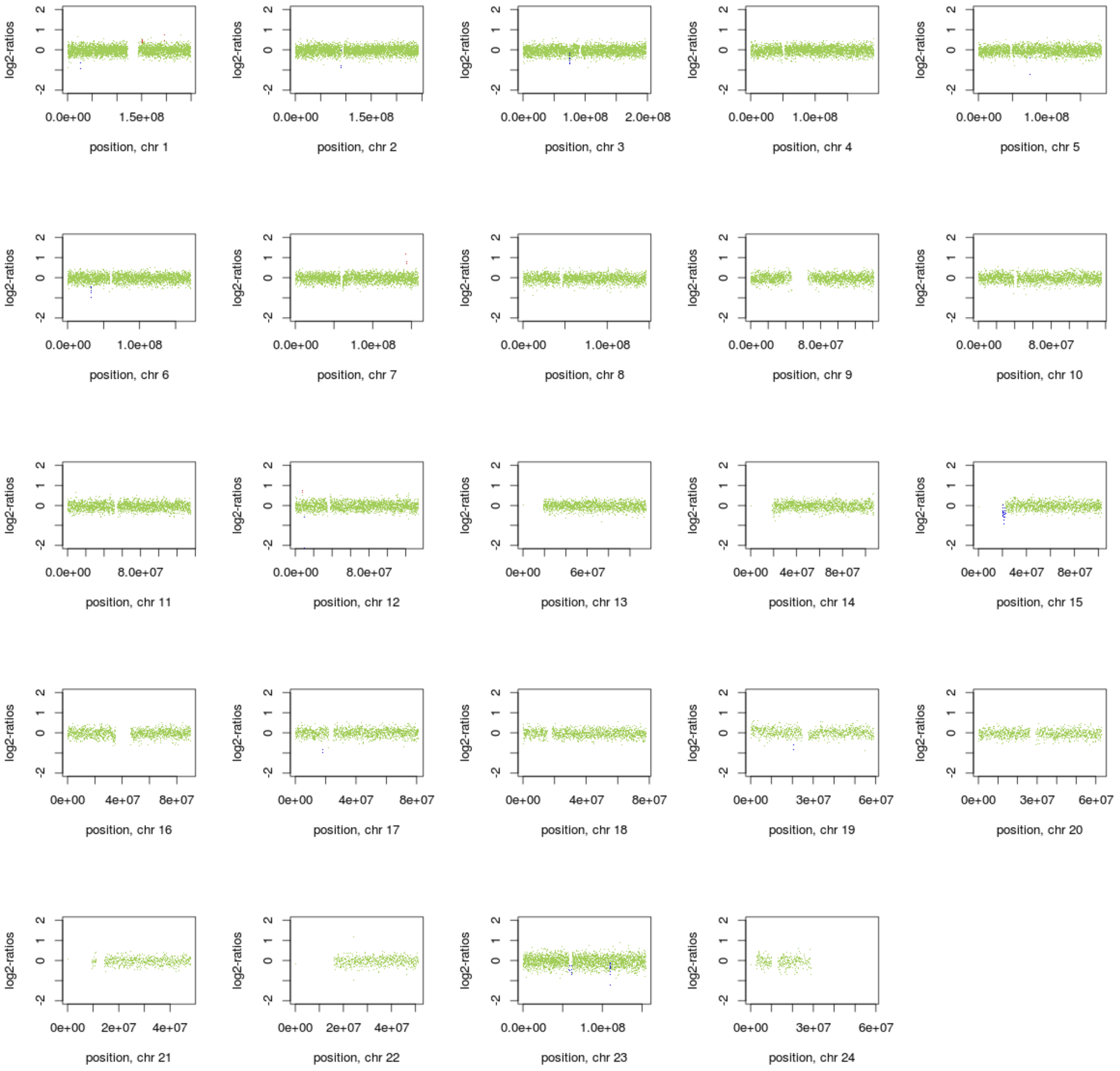


Figure 49: CNV profile for sample *Lunge1.2*

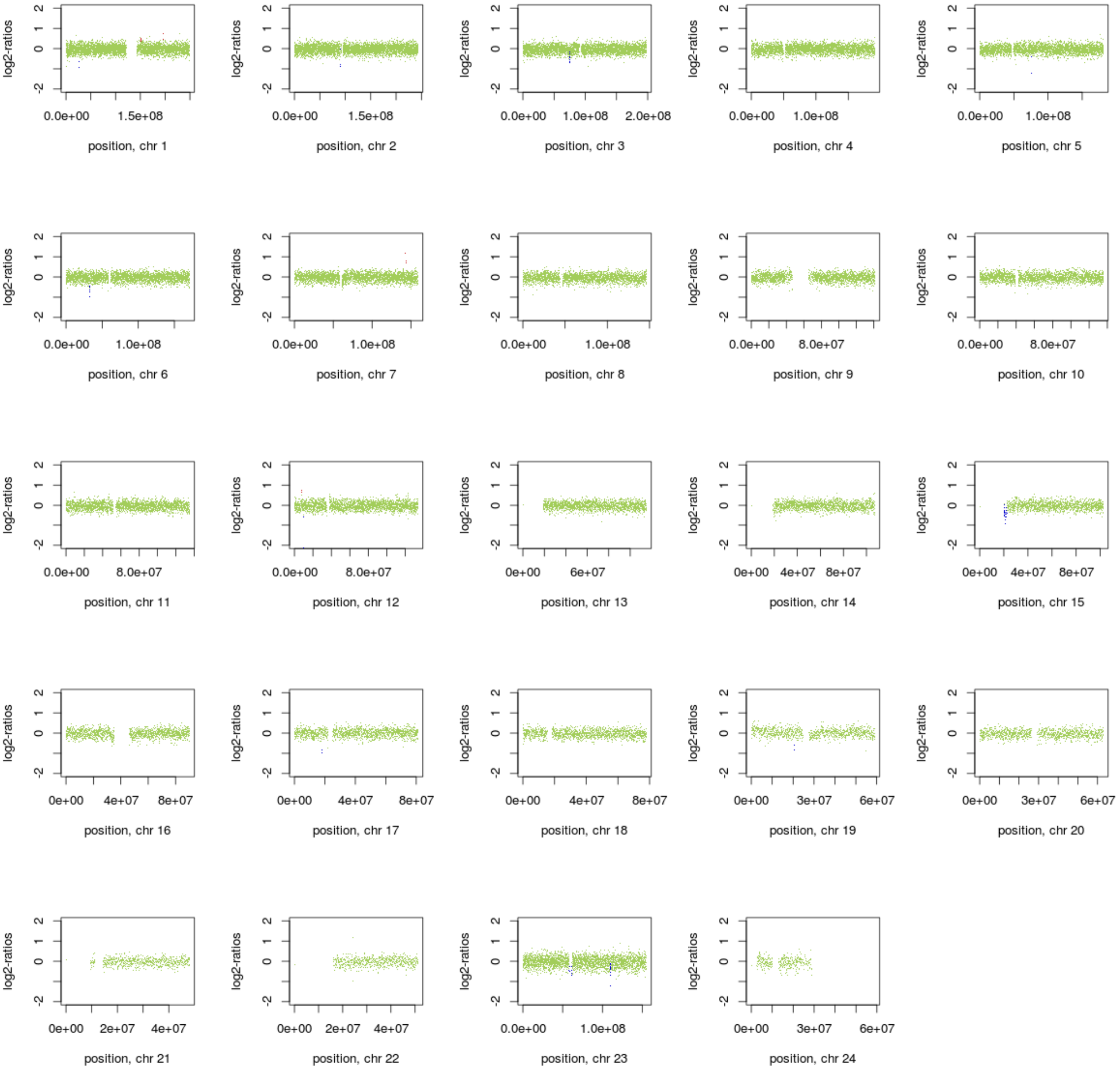


Figure 50: CNV profile for sample *Lunge1.3*

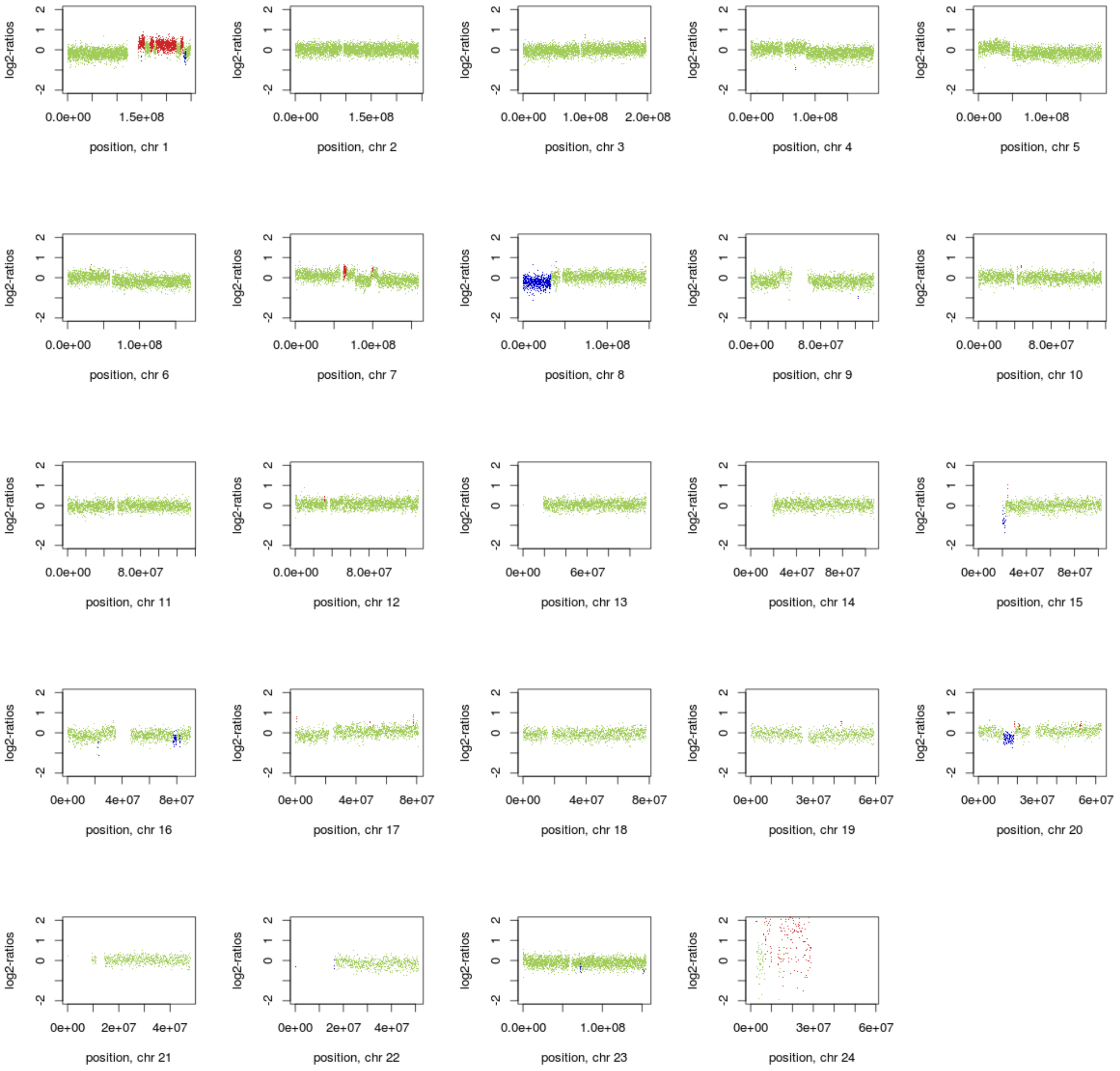


Figure 51: CNV profile for sample *Lunge2*

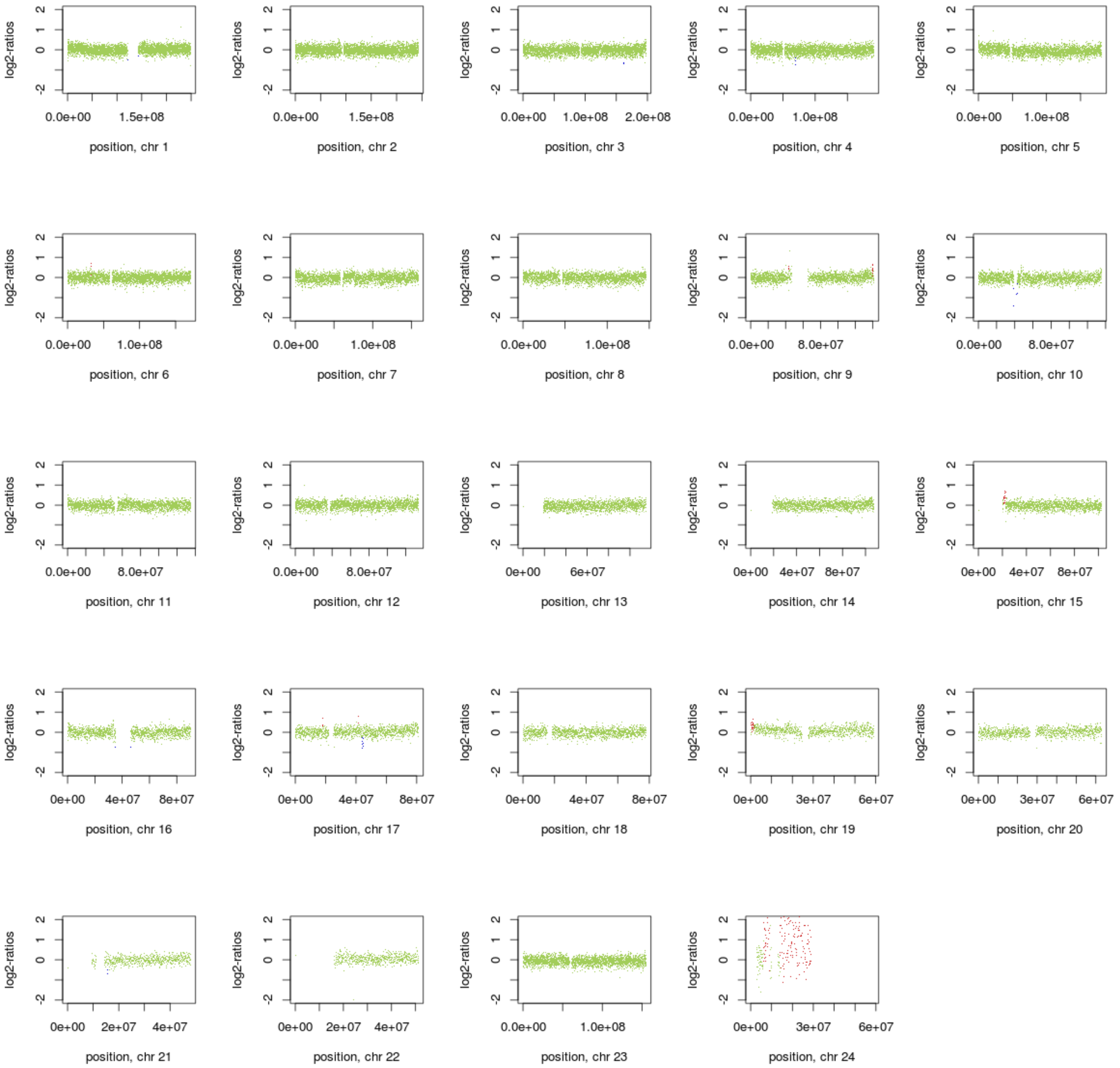


Figure 52: CNV profile for sample *Lunge3*

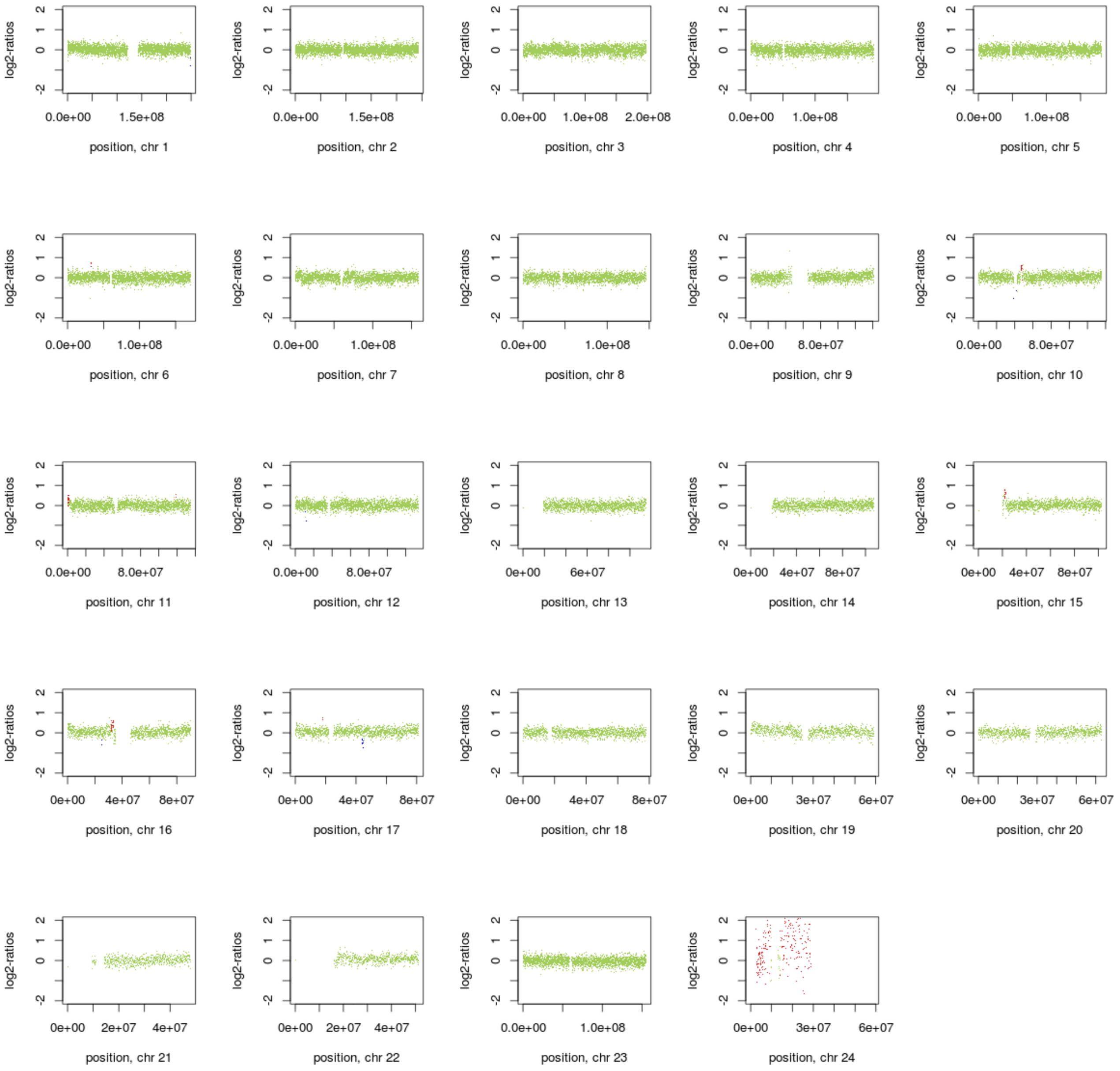


Figure 53: CNV profile for sample *Lunge3.2*

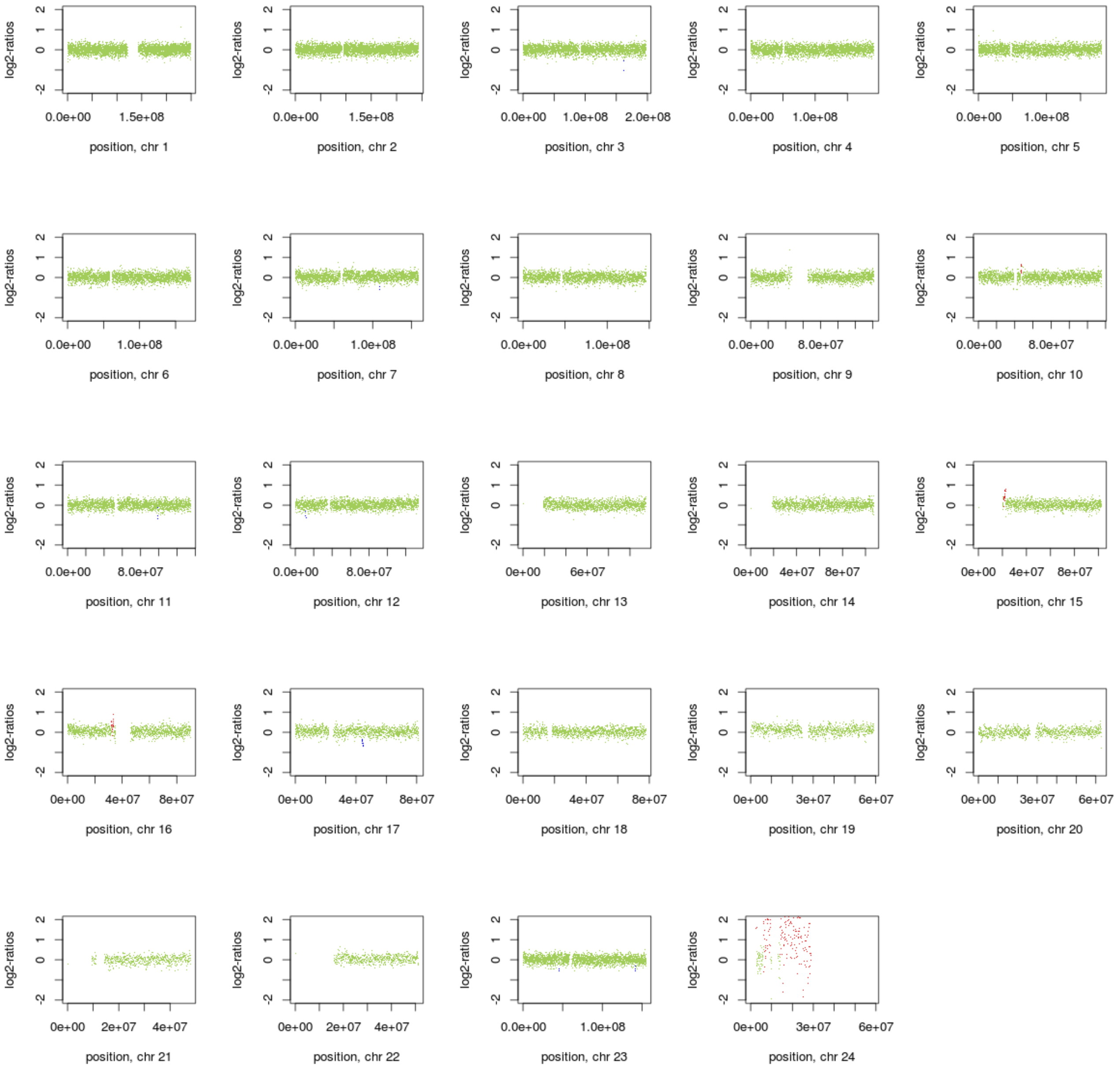


Figure 54: CNV profile for sample *Lunge3.4*

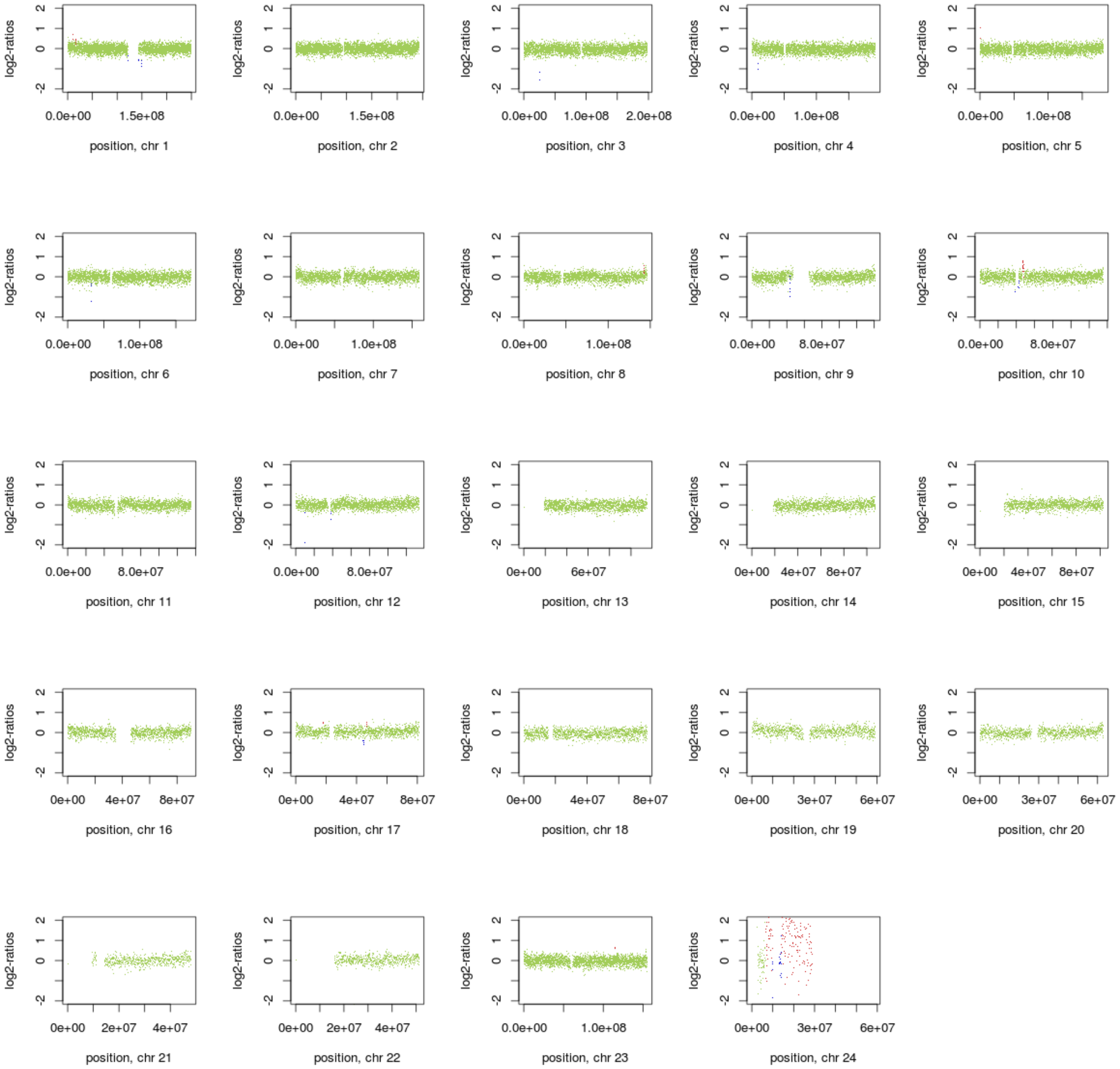


Figure 55: CNV profile for sample *Lunge4*

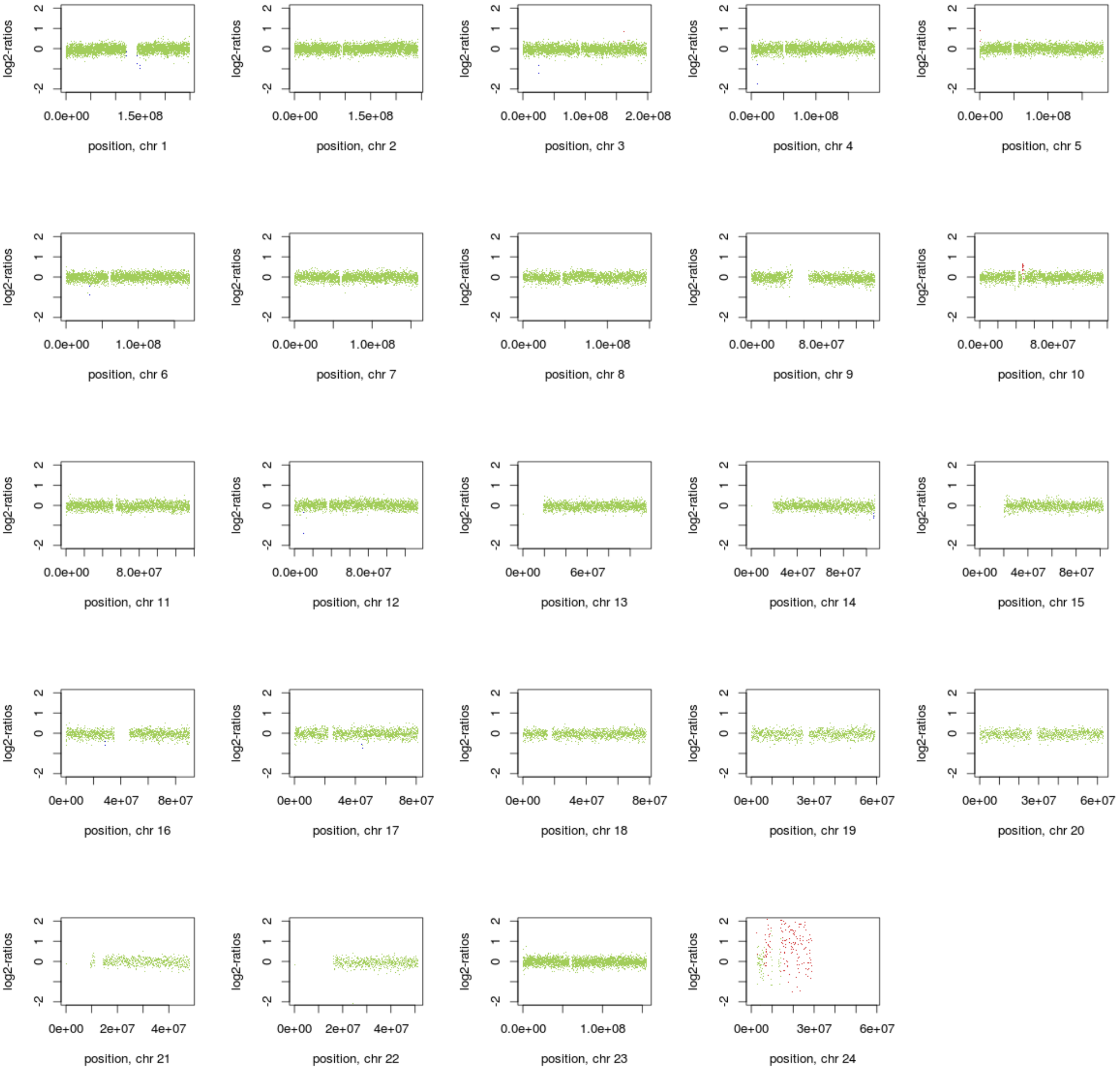


Figure 56: CNV profile for sample *Lung4.2*

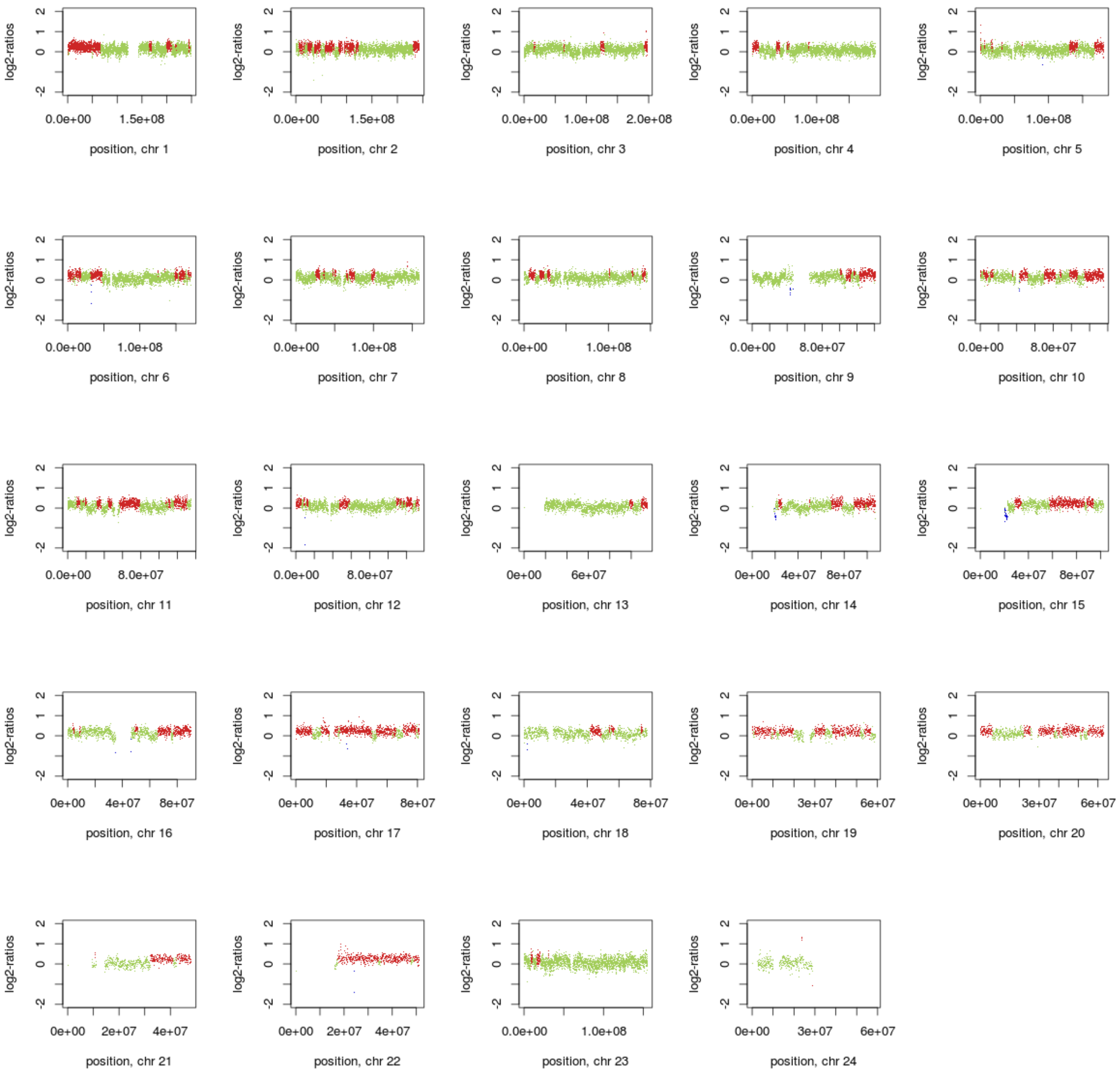


Figure 57: CNV profile for sample *Lunge5*

Segmental z-scores

In table 20 all segmental z-score lower or higher than +/- 3 for all samples are listed:

Table 20: Segmental z-scores < and > +/-3 for lung samples

Sample	Chromosome	Region start	Region end	z-score
Lunge 1	chr1 (p34.1)	45233897	45346789	28,17
	chr4 (q34.1)	175449608	175562341	-3,62
	chr4 (q34.3)	179181884	180028524	3,25
	chr7 (p14.3)	31935287	32503776	-3,83
	chr7 (q31.1)	108540129	108938011	-4,64
	chr7 (q31.2)	116316181	116431209	10,6
	chr7 (q32.1)	128759508	128872194	28,02
	chr7 (q35)	143239405	143477191	5,7
	chr7 (q35)	143477192	143571691	4,58
	chr9 (q22.32)	98167890	98280719	25,18
	chr10 (q24.32)	104234406	104403453	11,31
	chr16 (p13.3)	2051153	2164550	28,74
	chr17 (q11.2)	29528490	29698610	12,1
Lunge 1.1	chr3 (q29)	195812923	196829530	3,33
	chr5 (q13.3)	75794513	75963576	-3,28
	chr6 (q27)	170087734	170939734	-3,75
	chr7 (p22.1)	4782029	5628849	3,93
	chr7 (q35)	143239405	143571691	6,15
	chr11 (p15.4)	9255445	9368315	5,4
	chr12 (p13.33)	854981	1193628	3,36
	chr12 (p13.31)	7809241	8149872	3,82
	chr12 (q11-q12)	38001623	38512027	6,31
	chr12 (q24.12 - q24.13)	111933812	113122523	3,25
Lunge 1.2	chr1 (q21.3)	150552163	150779051	3,93
	chr1 (q31.3)	196818434	196933627	4,8
	chr4 (q13.1)	63450793	64359703	-3,36
	chr7 (q35)	143239405	143571691	5,33
	chr8 (q24.22)	135157765	135666453	-3,06
	chr11 (p11.2- q12.2)	47849357	61360693	-3,32
	chr12 (p13.31)	7979111	8092829	5,61
	chr12 (q11- q12)	38058342	38512027	3,64
	chrX (q23)	110031378	110540775	-4,32
Lunge 1.3	chr2 (p21)	42563495	42902196	3,2
	chr4 (q32.1 - q32.2)	161404337	161914213	-7,04
	chr5 (p14.1 - p13.3)	27291555	29102118	-3,23
	chr6 (p21.1)	42474620	42926203	3,65
	chr7 (q35)	143239405	143571691	5,75
	chr7 (q36.1- q36.2)	152158200	152673368	3,32
	chr10 (q11.22- q11.23)	49877405	51068010	-3,01
	chr10 (q23.1- q23.2)	87186926	88148210	-3,11

	chr11 (p15.4)	9255445	9368315	4,81
	chr12 (p13.31)	7979111	8149872	5,19
	chr12 (q11- q12)	38058342	38568378	7,81
	chr19 (p13.3)	738015	1132761	3,4
Lunge 2	chr1 (p36.33 - p35.3)	0	28071110	-3,95
	chr1 (p35.3-q21.1)	29538926	142786713	-11,04
	chr1 (q21.1)	145030632	145933216	6,41
	chr1 (q21.1-q21.2)	145933217	149043990	4,6
	chr1 (q21.2)	149932196	150101252	4,71
	chr1 (q21.2-q21.3)	150101253	150383107	5,48
	chr1 (q21.3)	150383108	151682642	9,65
	chr1 (q21.3)	151682643	153782041	8,77
	chr1 (q21.3)	153782042	154176888	10,23
	chr1 (q21.3-q22)	154176889	156162987	6,17
	chr1 (q23.3)	160917769	161637242	4,15
	chr1 (q24.2)	167357076	168374514	6,49
	chr1 (q24.2- q24.3)	168374515	171430858	4,11
	chr1 (q24.3- q25.1)	171430859	174203796	6,81
	chr1 (q25.2-q32.1)	179316048	206788069	14,32
	chr1 (q32.2- q41)	207295243	219697474	9,27
	chr1 (q42.13- q42.2)	228734504	234127836	8,54
	chr1 (q43)	237006902	237236963	-6,95
	chr1 (q43)	237236964	237861272	-5,15
	chr1 (q43)	237861273	239666402	-8,26
	chr1 (q43- q44)	239666403	245121864	-12,22
	chr2 (q22.3)	145692345	145974087	-3,55
	chr3 (p26.3 - p21.31)	229525	47145863	-3,17
	chr3 (p13)	73251982	73985627	-3,1
	chr3 (q12.2)	100316192	100429366	6,76
	chr4 (q13.2 - q21.23)	69519080	85564975	5,1
	chr4 (q21.23 - q22.1)	85564976	90780761	-4,31
	chr4 (q22.1 -q35.2)	92996387	190955048	-9,54
	chr5 (p15.1 - p13.3)	15590491	32439411	13,15
	chr5 (p13.3 -p11)	32439412	46349499	3,23
	chr5 (p11 -q35.3)	46349500	180730344	-9,22
	chr6 (q11.1- q11.2)	62180133	63368493	-7,78
	chr6 (q12)	64215839	65853800	-3,56
	chr6 (q12 - q26)	66650360	161831096	-9,6
	chr6 (q26- q27)	162394897	170939734	-7,17
	chr7 (p22.3- p21.1)	0	18687900	7,56
	chr7 (p21.1 -p12.1)	18687901	51525121	3,68
	chr7 (p12.1 -q11.1)	51525122	61088129	5,88
	chr7 (q11.21)	62471207	66123207	9,45
	chr7 (q11.21 - q11.22)	66123208	68503683	3,61
	chr7 (q11.22 - q11.23)	70312971	77346154	4,74
	chr7 (q11.23- q21.11)	77402497	79944338	-4,75
	chr7 (q21.11)	80000842	80847797	3,56
	chr7 (q21.11- q21.3)	80960482	97671630	-8,34

	chr7 (q21.3 - q22.1)	97727973	99593065	3,37
	chr7 (q22.1)	99593066	100503599	6,55
	chr7 (q22.1- q22.3)	100503600	106957804	3,41
	chr7 (q22.3 - q36.3)	107127375	159071691	-6,42
	chr8 (p23.3 - p12)	0	32961954	-7
	chr8 (p12)	32961955	33131711	-3,7
	chr8 (p12)	33131712	35452472	3,1
	chr8 (p11.22 -p11.21)	39407968	40598054	-4,87
	chr8 (p11.21 - q24.3)	40598055	146247376	3,32
	chr9 (p24.3- p24.2)	0	4350916	-5,7
	chr9 (p24.1- p21.1)	5143154	33042802	-10,3
	chr9 (p21.1 - p13.1)	33155845	39874183	4,44
	chr9 (q21.11 - q33.2)	70591881	122952464	-7,01
	chr9 (q33.2)	122952465	123065266	-8,17
	chr10 (q11.22)	47384970	47743030	7,45
	chr10 (q21.3 - q22.1)	69947095	71187931	4,58
	chr12 (p13.33)	0	2213006	4,22
	chr12 (p13.31 -p13.2)	5556064	11235137	3,7
	chr12 (p13.2)	11235138	11405865	-3,44
	chr12 (p13.2- p11.21)	11405866	31721752	3,33
	chr12 (p11.21)	31721753	32229798	5,84
	chr12 (p11.21 -q13.11)	32229799	47695055	4,62
	chr12 (q13.11- q13.13)	47695056	51875135	4,3
	chr12 (q13.13)	52557758	53237219	-3,11
	chr12 (q13.13- q24.31)	53237220	122146740	3,72
	chr12 (q24.31)	122146741	124343275	5,5
	chr14 (q22.3)	55947250	56454577	-3,29
	chr15 (q11.2)	24589511	24817962	4,29
	chr16 (p13.3 -p12.2)	3501850	22447369	-4,2
	chr16 (p12.2)	22447370	22757048	-6,85
	chr16 (p12.2 -p12.1)	22757049	24956934	-3,08
	chr16 (q23.1- q23.2)	77519267	79887542	-7,03
	chr16 (q23.2 - q23.3)	79887543	82033766	-3,22
	chr16 (q23.3)	82033767	82600048	-5,71
	chr17 (p13.3)	784586	954634	5,68
	chr17 (p13.3 -p12)	1124282	15048598	-3,74
	chr17 (q21.33)	49073690	49242867	5,62
	chr17 (q21.33-q22)	49242868	52634166	3,11
	chr17 (q22 - q23.2)	57266509	59106994	6,92
	chr17 (q25.3)	77619828	78132125	10,64
	chr18 (p11.31- q21.32)	3702784	58344609	-4,44
	chr18 (q22.1- q22.2)	64114190	68585224	-6,3
	chr19 (p11- q13.13)	24456884	35603172	-7,63
	chr19 (q13.2 -q13.31)	43308185	43709962	5,59
	chr20 (p13)	4473309	4698760	-5,83
	chr20 (p13- p12.2)	4698761	9339522	3,37
	chr20 (p12.1 -p11.23)	12852847	18158541	-8,94
	chr20 (p11.23)	18384416	18613189	5,24

	chr20 (p11.23)	20650911	21158088	6,25
	chr20 (p11.23- p11.22)	21158089	21553215	4,2
	chr20 (p11.1- q11.23)	25746742	36408008	3,38
	chr20 (q13.2)	51997923	52455454	4,94
	chr20 (q13.2- q13.31)	52455455	55229921	4,38
	chr20 (q13.33)	61820101	62895415	4,45
	chr22 (q11.22- q13.33)	22909025	51176570	-3,37
	chrX (p22.33 -q13.2)	3758059	72527442	-3,66
	chrX (q13.2)	72527443	73098078	-6,27
	chrX (q13.2 -q26)	73098079	149412962	-6,61
	chrX (q26)	149412963	151861608	-6,87
	chrX (q26)	151861609	152121295	-3,19
Lunge 3	chr1 (p36.33- p36.32)	936124	2481908	3,1
	chr1 (p11.2- q21.1)	121424191	142613953	-3,36
	chr2 (p11.2)	85514633	86362549	3,09
	chr3 (p25.3)	9007858	10305306	3,15
	chr3 (p24.3-p21.31)	20259538	46066750	-3,24
	chr4 (p14-q13.2)	40812649	69347419	-3,17
	chr5 (p12-q35.2)	44648750	173690920	-4,19
	chr6 (p24.3)	9055988	10019834	-3,4
	chr6 (p21.1)	42869835	43320736	3,12
	chr7 (p21.3- p11.2)	7542955	57915969	-3,15
	chr7 (q11.23)	72293922	76328348	3,08
	chr7 (q11.23- q21.3)	77458859	94550794	-4,12
	chr9 (q34.3)	139872428	140380614	3,55
	chr10 (q11.22)	46694097	47743030	3,05
	chr11 (q14.1-q22.3)	77911652	107235188	-3,92
	chr12 (q24.31)	121638835	122372398	5,42
	chr17 (q21.31)	41356845	41469907	6,76
	chr17 (q21.33- q22)	49695248	54672160	-3,51
	chr19 (p13.3)	341258	1529338	4,73
	chr19 (p13.3)	1529339	5533203	3,1
	chr19 (p12-q11)	24230544	28083586	-3,76
	chr19 (q13.41- q13.42)	53048635	53730030	4
	chr19 (q13.42)	53957693	55033930	3,4
	chr19 (q13.42)	55602670	56280255	3,45
	chr21 (q11.2)	15730587	15843550	-3,23
	chrX (p11.22- q26.2)	50080957	132165716	-4,44
	chrX (q26.3- q28)	136902573	152383792	-4,59
Lunge 3.2	chr1 (p21.1)	103356718	104306546	-3,17
	chr1 (p13.1-q21.1)	116682832	143469112	-3,43
	chr1 (q44)	248659140	248797040	-3,07
	chr2 (p24.1)	21053361	22831058	-4,24
	chr2 (p12-p11.2)	76909830	84497541	-3,97
	chr3 (p26.3-p26.1)	0	8553148	-3,46
	chr3 (p25.1-p23)	15557969	31699338	-3,07
	chr3 (q26.1)	162345808	167103184	-5,24
	chr5 (q12.1-q12.3)	62780624	64033287	-3,18

	chr5 (q14.1-q31.1)	79928524	130634572	-3,04
	chr7 (q11.23)	72414347	76328348	4
	chr8 (q23.1- q23.3)	108996942	114595245	-5,39
	chr10 (q11.22)	46694097	47941657	4,48
	chr10 (q21.3)	64575345	66044340	3,42
	chr10 (q21.3)	66044341	69552644	-3,81
	chr11 (p15.5)	302884	1374779	3,24
	chr11 (p15.4)	3840393	8973021	-3,11
	chr11 (p11.2)	45760942	46720907	3,47
	chr11 (p11.12-p11)	49450722	54934099	-6,48
	chr11 (q14.3)	89853447	92405196	-5,84
	chr11 (q23.3)	119002979	119115680	3,97
	chr12 (p13.2)	11805091	12596496	3,49
	chr12 (p11.22- p11.21)	30527991	32342736	3,06
	chr12 (q13.12- q13.2)	49505698	54930635	3,01
	chr12 (q21.31)	83653434	85015215	-6,63
	chr13 (q14.3-q31.1)	53727736	86560327	-3,2
	chr15 (q15.1)	41135118	42432138	3,61
	chr15 (q21.1)	46223454	46393112	-3,15
	chr15 (q26.1)	90273383	91684501	3,16
	chr16 (p12.1)	25238764	25352192	-4,9
	chr16 (p11.2)	28285319	28633747	3,79
	chr17 (p13.3)	558653	2877066	3,26
	chr18 (q12.3)	37768295	41503207	-4,2
	chr19 (p13.3)	0	5589624	3,5
Lunge 3.4	chr1 (p36.13)	16902184	17269135	5,19
	chr1 (p36.12- p36.11)	23874264	24383233	3,44
	chr1 (p35.3)	28580070	29087798	10,17
	chr1 (p35.1)	32683347	33699977	3,29
	chr1 (p34.3)	35400643	36078342	4,74
	chr1 (p13.3)	109262906	109826970	4,18
	chr1 (p13.3-p13.2)	111695449	112091567	-4,22
	chr1 (q21.2-q21.3)	149799349	151568115	4,03
	chr2 (p13.2)	71685266	71910842	-3,02
	chr2 (q21.3-q22.3)	136307878	146143133	-3,53
	chr2 (q33.1-q33.2)	203064959	204475788	5,94
	chr7 (p22.2- p22.1)	4093548	7144301	3,18
	chr7 (p15.3)	23102097	23610555	4,26
	chr7 (q11.21- q11.23)	62471207	76202880	3,67
	chr7 (q21.3- q22.1)	97840656	102105393	3,01
	chr10 (q11.22)	46694097	47675520	3,68
	chr11 (q14.1)	80794797	82435556	-3,86
	chr11 (q22.1)	98684230	98853714	-3,69
	chr11 (q22.3)	104353393	105202167	-4,51
	chr11 (q22.3)	107351539	107576944	3,98
	chr11 (q22.3)	107802553	108483499	3,53
	chr12 (p11.21- p11.1)	31205898	34492124	4,04
	chr12 (q12)	42481218	42707587	4,45

	chr12 (q13.12)	50521287	50915821	6,59
	chr12 (q13.13)	53011583	54478414	3,01
	chr12 (q13.2-q13.3)	56078096	57094132	5,17
	chr14 (q32.31- q32.33)	102290139	104376097	4,8
	chr19 (q13.41- q13.42)	52258439	54750038	5,5
	chr19 (q13.42-q13.43)	55432719	59054950	3,73
	chr20 (p13)	3001619	3283609	3,93
	chr20 (p12.3)	6961019	7530016	-4,13
	chr21 (q21.3)	30409251	31199391	-4,5
	chrX (p11.3)	45974297	46087025	-3,77
	chrx (p11.3-p11.23)	46087026	49455884	3,27
Lunge 4	chr1 (p36.13)	16902184	17326443	5,44
	chr2 (q37.1)	231563683	232409882	3,46
	chr3 (p24.3-p24.2)	16799705	26033776	-3,34
	chr3 (p24.2)	26033777	26146634	-5,14
	chr3 (p14.3 - q21.2)	58507155	124397806	-3,93
	chr3 (q22.1-q26.2)	130338621	168348647	-3,57
	chr5 (p15.33)	639902	761399	3,83
	chr5 (p13.1)	38869372	39602920	-4,41
	chr7 (p22.2- p22.1)	4150290	6611813	3,1
	chr8 (q24.3)	142116406	142512619	4,36
	chr9 (p24.3- p13.2)	1924446	38241246	-3,25
	chr10 (q11.22)	46694097	48185846	4,81
	chr10 (q26.3)	134920077	135203131	3,9
	chr11 (p11.2- q11)	48642399	55331032	-6,91
	chr11 (q23.3- q24.1)	120920624	121260123	-5,91
	chr12 (q11)	38001623	38171948	-3,34
	chr12 (q12)	40098986	40499087	-4,02
	chr12 (q24.23- q24.31)	120228943	121131329	4,14
	chr14 (p13-q21.3)	0	50127546	-3,04
	chr17 (q21.32)	46536092	46817880	8,47
	chr17 (q22- q23.1)	57210151	58065482	3,45
	chr19 (p13.3)	341258	5476714	3,6
	chr19 (p12)	21512233	22080006	3,38
	chr19 (p12-q11)	22080007	28083586	-3,96
	chr19 (q13.2)	38713587	39164538	3,93
	chr19 (q13.31)	44452388	44621460	-4,83
	chr21 (q22.3)	45122902	47096850	3,27
	chrX (p11.22- q23)	54350347	114905598	-3,17
	chrX (q23)	114905599	115056814	3,46
Lunge 4.2	chr3 (p24.2)	26033777	26146634	-4,29
	chr3 (q26.1-q26.2)	162627563	168292302	-3,43
	chr3 (q26.2)	168292303	170726480	4,85
	chr5 (p15.33)	639902	761399	3,01
	chr8 (q12.1-q21.13)	56625878	83581475	3,21
	chr10 (q11.22)	46694097	48185846	4,36
	chr10 (q22.1)	72147009	73501233	-3,46
	chr11 (q23.3)	116969262	117986834	-3,11

	chr15 (q14)	34378167	38140536	3,05
	chr16 (p11.2)	28633748	28884915	-3,29
	chrX (p21.1)	35342233	36813361	-3,31

In table 21 all segmental z-score lower or higher than +/- 3 for all controls are listed.

Table 21: z-scores < and > +/-3 for all controls

Sample	Chromosome	Region start	Region end	z-score
F19	chr1 (p31.1)	72700092	72812838	3,01
	chr1 (p22.2)	89437049	90117748	-3,93
	chr1 (p11.2-q12)	121340705	142552361	-4,06
	chr7 (q21.11)	85711581	86052909	-3,93
	chr7 (q22.1)	101362991	101588408	4,17
	chr7 (q35)	143176440	143797419	-3,86
	chr9 (p11.2)	44762055	44882887	-3,07
	chr10 (p11.1)	38700099	38987055	-3,27
	chr10 (q11.22)	46694097	47556794	3,50
	chr10 (q11.22)	47556795	47743030	14,80
	chr11 (p15.2)	14565471	14854157	-4,24
	chr11 (p11.2- p11.12)	48698808	48869153	-3,17
	chr11 (p11.12-q12.2)	48869154	60887941	-3,25
	chr12 (p12.2)	21018998	21188138	-4,93
	chr15 (q26.1)	91119981	91514902	3,73
	chr16 (p11.2)	31498032	31668605	-7,09
	chr16 (q23.3-q24.1)	84183825	85482739	3,88
	chrX (p22.2)	9747898	10424982	3,04
	chrX (p21.1)	36245368	36813361	-3,38
	chrX (p11.21- q26.2)	55320848	132504599	-3,35
M18	chr2 (p16.1- p15)	61001361	62750154	5,68
	chr2 (q22.3)	148457252	148573246	-4,07
	chr2 (q33.1- q33.2)	201310852	204644878	3,21
	chr5 (q13.2)	68850410	70615062	3,98
	chr9 (p13.3)	33779626	34174446	3,50
	chr10 (q11.21)	43895150	44121210	3,01
	chrX (p22.33)	0	2677955	-6,45
	chrX (q21.2- q21.33)	82242066	94086685	-3,13
	chrX (q28)	154970998	155197903	-4,60
	chrY (p11.32- p11.2)	0	6025587	-10,96
	chrY (p11.2)	6025588	6738337	-11,42
	chrY (p11.2)	6738338	9902394	-11,55
	chrY (q11.21)	13609809	13666338	-5,42
M19	chr4 (p16.1- p15.33)	10017596	13421542	-3,68
	chr4 (p14)	39401136	40078130	4,04
	chr6 (q16.1- q16.2)	93687967	99790764	-4,42
	chr10 (q25.1)	106943710	107056407	-7,58
	chr14 (q23.1- q23.2)	61728937	62123529	-6,31
	chr14 (q23.2)	62518337	63143819	-7,16
	chr14 (q24.3)	77552508	77665185	-3,28

	chr17 (q24.3)	69932935	70722361	-3,31
M20	chr2 (p24.3- p24.2)	16396458	16966221	-3,07
	chr2 (p23.1)	30795471	30909415	-8,53
	chr3 (q29)	195756578	196433523	3,63
	chr10 (q26.13)	124322914	124563567	-3,33
	chr10 (q26.3)	134127952	134523543	3,84
	chr16 (p13.12- p13.11)	14621835	15129076	3,88
M21	chr1 (p36.33)	0	764926	12,56
	chr1 (p32.2)	58588248	58762459	5,14
	chr2 (p22.3)	33060181	33176792	20,90
	chr2 (p16.1- p15)	60945012	62185345	4,01
	chr2 (p11.2)	89550597	89888947	115,76
	chr2 (p11.2- q21.2)	89888948	132958518	6,50
	chr2 (q21.2)	132958519	133072975	139,89
	chr2 (q31.1)	174343382	174630084	-3,55
	chr3 (p12.3)	75573751	75856566	8,83
	chr3 (p11.1 -q11.1)	90296406	93528338	13,09
	chr4 (p11)	49073140	49196084	150,51
	chr4 (q35.2)	190782009	190955048	7,54
	chr5 (p15.2- p15.1)	14910385	16327754	-3,39
	chr5 (p11- q11.1)	46292790	49591814	4,20
	chr5 (q13.2)	68850410	69427707	7,10
	chr5 (q13.3)	75794513	75907239	-3,21
	chr7 (p11.2)	57497018	57609972	15,79
	chr7 (p11.2)	57915970	57972489	18,60
	chr7 (p11.2- q11.1)	57972490	61088129	62,86
	chr7 (q11.1- q11.21)	61088130	61792359	9,10
	chr7 (q11.21)	61792360	61854612	4,71
	chr7 (q11.23)	76068608	76638258	11,15
	chr7 (q36.1)	151932230	152158199	5,83
	chr8 (p11.22- p11.1)	39351614	43707260	37,95
	chr9 (p13.1)	38820140	38925146	3,34
	chr9 (p13.1)	38925147	39765747	4,54
	chr9 (p12)	41813190	43128289	3,65
	chr9 (q13)	66766668	66981945	153,26
	chr9 (q13- q21.11)	66981946	70430813	18,06
	chr10 (p11.1)	38763615	38928870	41,34
	chr10 (p11.1- q11.21)	38928871	42644039	5,91
	chr10 (q11.21)	42644040	42818391	33,55
	chr11 (p11.2- p11.12)	48755467	48993836	18,19
	chr11 (p11.12)	50609383	50779566	9,91
	chr11 (p11.12- q11)	51515741	55047332	27,25
	chr12 (p11.1- q11)	34831610	38001622	14,68
	chr12 (q11)	38001623	38171948	27,35
	chr12 (q11- q12)	38171949	38568378	3,95
	chr15 (q11.2)	22656224	23277294	12,02
	chr15 (q25.2)	82607709	83042495	4,33
	chr16 (p11.2)	33830540	34210300	40,00

	chr16 (p11.1- q11.2)	35235749	46456243	-3,21
	chr17 (p11.1- q11.1)	22210884	25328089	33,53
	chr17 (q21.32)	45176973	45289779	20,12
	chr18 (p11.32)	0	1725779	29,05
	chr18 (p11.32)	1725780	1838904	-4,78
	chr18 (p11.32- q23)	1838905	77960362	3,04
	chr19 (p11- q11)	24569574	27914292	23,04
	chr20 (p11.1)	25688707	26156310	3,06
	chr20 (p11.1- q11.21)	26156311	29429049	33,51
	chr20 (q11.21)	29429050	29863829	15,63
	chr21 (p11.2)	10594469	10760546	6,92
	chr21 (p11.2)	10760547	10873939	31,11
	chr22 (q11.21)	18619391	18941453	17,05
	chrX (q11.1)	61723346	61950448	41,31
	chrX (q11.21)	61950449	155197903	-3,01
	chrY (p11.2)	9902395	10071873	41,45
	chrY (p11.2- q11.21)	10071874	13609808	15,16
	chrY (q11.21)	13609809	13742968	12,03
	chrY (q11.21)	13742969	13910430	148,94
	chrY (q11.223)	24591325	25543797	11,91
	chrY (q11.23)	26518424	27025638	-11,20
M30	chr1 (q31.3)	196697091	196818433	-3,35
	chr2 (p23.1- p22.3)	31930922	32834643	3,95
	chr10 (q11.21)	42701742	42760445	-3,25
	chr13 (q13.1)	33363315	33476975	-6,33
	chr15 (q14)	34659822	34860585	-6,71
	chr15 (q21.2)	50636554	51427777	3,22
	chr16 (p13.11)	16573881	16717239	-3,63
	chr16 (p12.3)	18528574	18755723	-4,09
	chr20 (p11.23)	18328025	19804551	-3,73
M31	chr2 (p11.2- p11.1)	89832448	91674412	-4,51
	chr2 (q11.2- q12.1)	100491800	103092377	-3,79
	chr2 (q12.1- q12.3)	105240946	108855433	-3,24
	chr6 (p21.32- q27)	32652485	170939734	-3,26
	chr8 (q24.23)	137652268	137935571	-5,38
	chr8 (q24.3)	142003677	142455902	3,00
	chr8 (q24.3)	145738191	146247376	-3,02
	chr12 (q23.3)	107596746	108781283	-3,13
	chr13 (q12.3)	31215273	31723136	-3,73
	chr13 (q13.1)	32855602	32968350	6,98
	chr15 (q12)	26735245	26961032	-3,07
	chr16 (p13.3)	690114	972557	4,75
	chr16 (q12.1)	48610729	49121229	3,87
	chr16 (q12.1- q21)	49121230	66491689	-3,29
	chr17 (q21.31)	41187599	41300373	8,31
	chr19 (p11- q12)	24513240	31932470	-3,56
	chr20 (q13.13)	47092013	47374997	-3,41
	chr22 (q12.3)	34241783	35317296	-3,37

	chrX (q25)	120980631	121207178	6,54
M32	chr6 (p21.32- q27)	32652485	170939734	-3,11
	chr8 (q23.3 -q24.13)	116638691	125408274	-3,79
	chr9 (q31.2)	108609701	109063405	-3,21
	chr12 (p13.33)	854981	1193628	3,14
	chr12 (q14.1- q15)	61396196	68985601	-3,44
	chr13 (q21.1)	55534456	55816347	3,52
	chr14 (q12)	25190798	27232252	-3,44
	chr15 (q14)	34659822	34860585	-5,07
	chr15 (q21.1)	46167126	47243053	-4,51
	chr17 (q25.3)	79656290	79932144	3,61
	chr19 (q13.42)	55997554	56222941	3,45
	chr21 (q22.13- q22.3)	39151552	43944059	-3,31
M33	chr2 (p11.2)	87366691	88008829	4,83
	chr8 (p12)	30474008	30699536	4,88
	chr9 (p13.3)	33836080	34401571	3,70
	chr20 (q13.2)	52173316	52512073	3,95
M34	chr2 (p16.1- p15)	60888652	62806506	5,23
	chr2 (p12)	77025207	82515192	-4,69
	chr2 (q33.1- q33.2)	201707774	203573124	3,10
	chr7 (q32.2)	129549370	129774753	3,88
	chr10 (p11.21)	35060882	35568356	3,33
	chr11 (q11)	55331033	55444133	-3,07
	chr14 (q24.3)	73827699	74168946	3,45
	chr17 (p11.2)	18890321	19080907	-3,20
	chr18 (p11.21- q11.1)	13223177	18647088	-3,26
	chr18 (q11.2- q12.1)	24922731	27811111	-3,45
	chr18 (q21.1)	43536470	43818491	3,59
	chr19 (p13.3- p13.2)	6887190	7000051	4,55
	chr21 (q21.3)	29619431	29788402	-4,18
F2	chr1 (p36.23- p36.22)	8177336	10658327	3,57
	chr1 (p36.11- p35.3)	26998281	29369758	3,00
	chr1 (p35.2- p35.1)	32288590	33530762	3,07
	chr1 (q31.3)	196757061	196875068	-3,40
	chr2 (p11.2)	87443224	88008829	3,33
	chr3 (p14.3- p14.2)	58507155	60095990	-3,10
	chr3 (q12.3)	101277799	101562514	4,60
	chr3 (q26.2)	169876604	170045793	5,81
	chr5 (p13.3)	32100307	32214002	6,68
	chr5 (q31.2- q31.3)	138443328	140355787	3,48
	chr7 (q36.3)	157655011	158505313	-3,14
	chr9 (p21.2)	26307397	26420284	-3,46
	chr9 (p13.1)	39331682	39595373	-4,85
	chr9 (q33.3)	129676812	129959792	-5,68
	chr10 (q21.3- q22.1)	70059792	70680417	6,37
	chr10 (q23.1- q23.2)	86899171	88148210	-5,85
	chr11 (p15.4)	9199094	9595223	3,14
	chr11 (p11.2- q12.1)	48357767	59805216	-4,30

	chr12 (p11.21)	31265134	32963509	6,36
	chr12 (q13.12)	50521287	51197630	6,24
	chr12 (q24.23)	118931637	119044516	-8,82
	chr12 (q24.23- q24.31)	120624278	123948324	3,24
	chr13 (q12.11)	21324882	21719443	5,10
	chr14 (q13.1- q13.2)	34647516	35785302	4,53
	chr14 (q32.31- q32.32)	102120416	103417911	3,72
	chr15 (q15.1)	41304165	41925016	3,35
	chr15 (q25.3)	86818280	87102434	-9,00
	chr17 (p13.3)	1974002	2933528	3,55
	chr17 (q21.33- q22)	49412452	54897618	-3,75
	chr19 (p12)	21003681	21117319	-3,89
	chr19 (p12)	21286699	21909354	5,99
	chr19 (q12)	29555355	29895341	-4,35
	chr19 (q13.31)	43709963	43824490	-3,73
	chr19 (q13.32- q13.33)	47613213	51128236	3,85
	chr20 (q12)	40252998	41667132	-4,71
	chr20 (q13.33)	61042922	61205749	3,19
F3	chr1 (p36.13)	16781783	17269135	4,35
	chr1 (q44)	248247343	248479118	6,52
	chr3 (q13.32)	117714758	117884394	6,96
	chr3 (q22.3)	135850484	136363147	3,62
	chr4 (p11- q11)	49551925	52685223	-3,09
	chr5 (p15.31- p15.2)	7789963	13044803	-3,58
	chr7 (q22.1)	99593066	100503599	3,42
	chr8 (q23.3- q24.3)	114709527	144501355	-3,09
	chr15 (q26.2)	96708904	96821611	-3,26
	chr21 (q21.1)	16749088	17144131	-3,76
	chr21 (q21.3)	29336855	30127286	-4,26
	chrX (q22.1- q23)	101417388	113600079	-4,87
F4	chr1 (p36.13)	16781783	18848245	-3,71
	chr1 (p33)	48185508	49660622	-5,37
	chr6 (p21.2- p21.1)	39822141	40612014	-3,05
	chr9 (p24.1)	5086806	5199641	14,64
	chr9 (q34.11)	132552843	132834741	4,71
	chr11 (p15.4- p11.2)	9595224	47172869	-3,03
	chr11 (p11.2)	47172870	48074805	3,73
	chr11 (p11.2- q12.1)	48074806	56806502	-3,40
	chr11 (q12.1)	56806503	57144895	-4,95
	chr11 (q13.3- q25)	68765883	134889487	-3,23
	chr17 (q21.31)	44392869	44769626	3,11
	chr17 (q22)	51957039	54389876	-4,28
	chr19 (p13.3)	3277297	4518241	3,10
F5	chr1 (p36.11- p35.3)	26714709	29482489	3,63
	chr1 (q21.3)	152421991	153329735	-6,18
	chr2 (p11.2- p11.1)	89832448	91674412	-4,26
	chr2 (q14.1- q14.3)	117254269	125302242	-5,43
	chr3 (p14.3)	56588214	57943633	5,41

	chr3 (q12.2)	100316192	100429366	8,04
	chr4 (q13.2)	69462589	70142292	-3,23
	chr4 (q13.2)	70142293	70256442	-3,81
	chr7 (p22.1)	5234164	7031251	3,91
	chr7 (q31.31)	118409417	120218446	-5,14
	chr9 (p11.2)	44762055	44882887	-3,33
	chr11 (p14.1- p13)	29008362	32406428	-5,71
	chr11 (p11.2)	46946460	48244141	4,19
	chr11 (q22.3)	107689647	108027955	4,87
	chr16 (p11.2)	29437593	31554471	3,58
	chr17 (q21.2)	39320976	39550271	-5,60
	chr17 (q22- q23.1)	56645726	58193524	4,62
	chr18 (q12.2- q12.3)	33864050	41164136	-5,74
	chr19 (p13.3)	794349	2036918	4,37
	chr19 (p13.3)	2770044	2939095	-4,92
	chr19 (p13.3- p13.12)	2939096	14901573	3,27
	chr19 (q13.31- q13.43)	43886999	59054950	3,55
	chr20 (p11.22- p11.21)	22009074	25179630	-3,18
	chr20 (q13.33)	58782897	59630739	-3,55
	chr20 (q13.33)	62328670	62724957	5,53
F6	chr1 (q25.2)	176427493	177161625	-6,01
	chr5 (q23.2)	124746238	124915665	-4,61
	chr5 (q32- q33.2)	146074417	154919058	-4,17
	chr6 (q13)	73958884	74241494	7,92
	chr7 (q22.1)	100895886	101701091	3,45
	chr9 (p23)	11873980	12099597	-3,65
	chr15 (q14)	37233437	39213557	-6,34
	chr20 (p12.2)	11202288	11427661	-5,51
	chr22 (q12.1)	27069836	28084461	-3,05
F7	chr1 (p36.33- q44)	0	249178441	-3,88
	chr2 (p25.3- q37.3)	0	243052394	-4,56
	chr3 (p26.3- q26.1)	0	162514872	-5,01
	chr3 (q26.1- q29)	162627563	197857588	-3,90
	chr4 (p16.3- p11)	0	48960083	-4,54
	chr4 (p11- q35.2)	49551925	190955048	-4,18
	chr5 (p15.33- q23.3)	0	127916966	-4,44
	chr5 (q23.3- q35.1)	127916967	171260129	-5,71
	chr6 (p25.3- p21.32)	0	32483026	-3,22
	chr6 (p21.32- q27)	32708978	170939734	-4,45
	chr7 (p22.1- p12.1)	6724500	52600040	-4,77
	chr7 (p12.1)	53448930	53561774	-6,68
	chr7 (p12.1- q36.3)	53561775	159071691	-3,21
	chr8 (p23.1- q24.21)	12500095	131307673	-4,73
	chr8 (q24.21- q24.23)	131307674	137652267	-5,35
	chr8 (q24.23)	137652268	137708903	-3,58
	chr8 (q24.23)	137708904	137878779	-7,90
	chr8 (q24.23- q24.3)	137878780	141438582	-5,80
	chr9 (p24.3- p13.1)	0	40805614	-5,09

	chr9 (p13.1- p12)	40805615	41920969	3,98
	chr9 (q21.12- q31.3)	72506599	111325440	-3,69
	chr9 (q31.3)	111325441	111495118	-5,58
	chr9 (q32)	115401503	115631934	-5,74
	chr9 (q32- q33.1)	115631935	121929708	-3,26
	chr9 (q33.1)	121929709	122269542	-6,41
	chr9 (q33.1- q33.2)	122269543	122779703	-3,08
	chr9 (q33.2)	122779704	123178122	-5,52
	chr10 (q11.22)	46584317	47287882	-3,40
	chr10 (q11.22)	48841348	49361722	-4,07
	chr10 (q11.22- q26.3)	49482713	135438600	-3,82
	chr11 (p15.5- q13.1)	0	64472231	-5,28
	chr11 (q13.3- q25)	68765883	134889487	-5,29
	chr12 (p13.33- q24.33)	0	133784555	-3,43
	chr13 (q12.13- q14.2)	27031705	48647134	-4,68
	chr13 (q14.2-q34)	48647135	115053231	-5,03
	chr14 (q11.2- q32.2)	22410264	99637711	-4,45
	chr17 (q11.2-q12)	30947586	33159344	-4,28
	chr18 (p11.32- q23)	0	77960362	-4,99
	chr19 (p12- q13.11)	23891009	32778882	-6,11
	chr20 (p13- q13.33)	0	61149406	-3,33
	chr21 (p13- q22.3)	0	48055817	-3,08
	chrX (p22.33-q28)	0	155197903	-5,11
F8	chr1 (p36.33- q21.2)	0	149043990	-3,85
	chr2 (p25.3- q37.3)	0	243052394	-3,48
	chr3 (p26.3- q29)	0	197857588	-3,63
	chr4 (p16.3- q35.2)	0	190955048	-4,05
	chr5 (p15.33- q35.3)	0	180730344	-3,50
	chr6 (p21.32- q27)	32652485	170939734	-3,59
	chr7 (q31.1)	112459715	112573337	-6,04
	chr8 (p23.1- q24.3)	8072633	144219320	-3,93
	chr9 (p24.3- p11.2)	0	44226814	-3,30
	chr9 (q21.11- q21.33)	70967798	87248854	-4,51
	chr9 (q31.1- q33.3)	106740878	127644896	-3,07
	chr10 (q11.22)	46694097	47743030	3,15
	chr11 (p15.5- q25)	0	134889487	-3,30
	chr13 (p13- q34)	0	115053231	-3,79
	chr16 (p11.1- q24.3)	34608574	90173456	-3,37
	chr17 (q25.3)	79035762	79543206	3,75
	chr18 (p11.32- q23)	0	77960362	-3,23
	chrX (p22.33- q11.2)	0	64511461	-3,24
	chrX (q11.2- q21.1)	64511462	81164426	-6,62
	chrX (q21.1- q28)	81164427	155197903	-5,15
F9	chr1 (p35.3- q21.2)	29651625	149043990	-3,42
	chr3 (p14.3- q13.32)	58564318	118448659	-4,81
	chr4 (p16.3- q13.2)	0	70142292	-3,01
	chr4 (q13.2)	70142293	70256442	-4,13
	chr5 (q23.1)	117329453	117781961	-9,58

	chr8 (q24.23)	137595912	137708903	-3,67
	chr9 (p21.3)	24328583	24441657	-3,23
	chr10 (q11.22)	47556795	47743030	8,98
	chr11 (p15.4- q13.1)	3313624	63625360	-3,19
	chr11 (q13.1)	64979470	65205247	4,08
	chr11 (q14.1- q23.3)	78137476	118550893	-3,69
	chr15 (q21.3)	54206520	55342036	-4,96
	chr15 (q25.3- q26.1)	86705611	89480896	-4,28
	chr17 (q24.2- q25.1)	66246187	71737068	-3,12
	chr20 (p12.3- q11.21)	5829841	29863829	-3,19
	chr22 (q11.23)	25656937	25714476	6,05
	chr22 (q11.23- q12.1)	25714477	25941946	13,87
	chrX (p11.22- q21.2)	54575782	85138178	-3,22
	chrX (q21.2- q28)	85138179	155197903	-3,35
F10	chr1 (p34.3- p34.2)	39293896	40655933	5,77
	chr1 (q41)	219983361	221736126	3,39
	chr1 (q42.12- q44)	225340040	249178441	3,64
	chr2 (p25.3- p25.1)	0	11650593	3,62
	chr2 (p24.2- p16.3)	17362110	51775416	3,27
	chr2 (q31.1)	170557761	175820129	3,97
	chr2 (q32.3- q36.3)	195588703	230095019	3,38
	chr3 (p22.3)	32602160	33053152	3,56
	chr3 (p21.2)	51383317	51665735	8,21
	chr3 (p21.2- p14.2)	51665736	58852781	4,26
	chr3 (q13.33- q24)	119922307	143665140	3,15
	chr5 (p15.1)	16893526	17119233	6,20
	chr5 (q33.2- q34)	155309379	160176230	3,76
	chr6 (q22.33- q23.2)	129706278	132251872	3,23
	chr6 (q25.2- q25.3)	153960283	160224654	3,17
	chr6 (q27)	165962877	170939734	3,11
	chr7 (p11.2)	55834956	56116956	5,08
	chr7 (q11.21- q21.11)	65269300	78192235	3,27
	chr7 (q21.3)	96929988	97386144	-3,49
	chr8 (p23.1- p11.22)	8298690	39238450	3,07
	chr9 (p24.1- p23)	8934288	13910460	-3,18
	chr9 (p21.1- p12)	31909476	43329397	3,30
	chr10 (p15.3- p11.1)	0	38763614	3,84
	chr10 (q21.1)	57624015	59546946	-3,52
	chr10 (q23.32- q24.33)	93539091	105306205	3,90
	chr11 (p11.2)	45986620	47962104	3,16
	chr12 (p13.31)	9642168	9762636	3,45
	chr12 (p13.2- p13.1)	12030473	13273244	3,62
	chr12 (p11.21)	31149230	32681617	6,77
	chr12 (q13.12)	50521287	51084957	5,92
	chr12 (q13.12- q13.13)	51084958	54083445	3,72
	chr12 (q21.33- q24.11)	91189072	109063118	3,75
	chr12 (q24.31)	122817750	123214776	4,37
	chr13 (q31.3- q32.1)	93207011	96886949	3,96

	chr13 (q32.2- q33.1)	98587531	103555559	3,12
	chr14 (q11.2)	20428725	24509190	3,00
	chr14 (q24.3)	73940686	74168946	5,16
	chr14 (q24.3- q31.1)	74168947	80207335	3,33
	chr14 (q31.3- q32.2)	88917322	97212877	3,06
	chr16 (p12.2)	22570823	22700218	-11,81
	chr16 (q22.1- q23.1)	67338734	75765901	3,13
	chr17 (p13.3)	0	2933528	3,32
	chr17 (p13.3- q21.2)	2933529	39888489	3,36
	chr17 (q21.2- q21.31)	39888490	44152556	3,41
	chr17 (q22- q24.3)	54728518	67210340	3,34
	chr18 (p11.31)	3364532	3702783	4,87
	chr18 (p11.31- p11.23)	6985585	7324275	20,21
	chr18 (p11.21)	12204146	12320735	6,19
	chr18 (p11.21)	12320736	12377079	5,64
	chr18 (q11.2)	23622122	23905198	11,78
	chr18 (q12.1)	26850103	27357785	-3,21
	chr18 (q12.1- q12.2)	27357786	34427714	3,42
	chr18 (q12.3- q21.31)	42124790	55118884	3,12
	chr18 (q21.31- q21.32)	55118885	57608479	4,01
	chr20 (p13- p12.3)	0	5773497	3,06
	chr20 (q13.12- q13.33)	45393680	58444509	3,04
	chr21 (q22.11- q22.3)	32327414	45291866	3,32
	chr22 (q11.21)	20296325	20660430	5,38
	chrX (p21.1- p11.21)	37204465	55969381	3,06
	chrX (q28)	153388870	153561789	5,35
F11	chr1 (p36.11- p35.3)	24946915	29651624	3,63
	chr1 (q21.2- q21.3)	150157603	151795481	3,09
	chr1 (q21.3- q22)	153499955	156162987	3,26
	chr2 (p25.3- p25.1)	0	9276579	-3,35
	chr2 (p25.1- p24.3)	11255456	12440354	3,16
	chr2 (p24.3- p24.1)	12440355	23738688	-6,67
	chr2 (p23.1)	30795471	31420726	-4,62
	chr2 (p23.1- p22.3)	31818161	32157257	-4,43
	chr2 (p16.3- p16.1)	48668381	60437540	-3,32
	chr2 (p14)	64728413	65124587	6,26
	chr4 (p16.3- p16.1)	3234936	6238619	-3,06
	chr4 (p14)	38497732	40755449	4,24
	chr7 (p22.3)	416748	1095648	3,42
	chr7 (p22.1)	5346893	6860038	4,41
	chr7 (q11.23- q21.11)	72236470	77515200	3,93
	chr7 (q21.11)	79606036	83164632	4,79
	chr7 (q22.1)	99423738	101475673	6,11
	chr7 (q32.1- q32.3)	128076823	130894124	3,54
	chr7 (q33)	133717694	133830456	-4,96
	chr8 (p23.1)	7994065	8072632	-3,32
	chr8 (q24.3)	141947150	142286316	4,16
	chr8 (q24.3)	144670583	145738190	3,59

	chr9 (q34.2)	137104487	137387092	4,86
	chr10 (q24.32- q24.33)	103839637	105757297	4,01
	chr10 (q26.13)	126026414	126815483	3,33
	chr11 (q12.3- q13.3)	63343296	68991488	3,86
	chr12 (p13.31- p13.2)	6120058	10554024	3,25
	chr12 (q13.2- q13.3)	56021642	57094132	6,03
	chr14 (q13.3- q21.3)	36861287	49957878	-4,88
	chr16 (p13.2)	9025617	9477670	3,41
	chr16 (p13.2)	9477671	10441990	-3,42
	chr16 (q21- q22.1)	66491690	67169411	3,88
	chr16 (q22.1)	68809749	68978963	-4,20
	chr16 (q22.1)	68978964	70477082	4,72
	chr17 (q22)	52182654	54897618	-3,75
	chr17 (q24.2)	65060895	66417158	3,54
	chr19 (p13.3)	341258	3841048	5,77
	chr19 (p12- q13.11)	22419147	34471586	-3,27
	chr19 (q13.31- q13.33)	43824491	51128236	3,05
	chr19 (q13.41- q13.43)	52031235	59054950	3,92
	chr20 (q13.33)	58670036	60082373	-3,63
	chr22 (q13.2- q13.31)	43454666	44300079	-3,27
	chr22 (q13.31)	46385349	47231447	4,84
	chrX (p11.23- q28)	49399505	155197903	-4,53

Index sequences

Table 22: Index sequences

Index	Sequence	Index	Sequence
1	ATCACG	13	AGTCAA
2	CGATGT	14	AGTTCC
3	TTAGGC	15	ATGTCA
4	TGACCA	16	CCGTCC
5	ACAGTG	17	GTAGAG
6	GCCAAT	18	GTCCGC
7	CAGATC	19	GTGAAA
8	ACTTGA	20	GTGGCC
9	GATCAG	21	GTTTCG
10	TAGCTT	22	CGTACG
11	GGCTAC	23	GAGTGG
12	CTTGTA	24	GGTAGC