

Master's Thesis

Implementation and Comparative Assessment of Diagnostic Cancer Gene Panels in the Molecular Pathology Laboratory

University of Luxembourg

Faculty of Science, Communication and Technology

Master in Integrated Systems Biology

by

Ben Flies

(010081174D)

Abstract

This is the template for a thesis at the Chair of Econometrics of Humboldt–Universität zu Berlin. A popular approach to write a thesis or a paper is the IMRAD method (Introduction, Methods, Results and Discussion). This approach is not mandatory! You can find more information about formal requirements in the booklet ‘Hinweise zur Gestaltung der äußeren Form von Diplomarbeiten’ which is available in the office of studies.

The abstract should not be longer than a paragraph of around 10 to 15 lines (or about 150 words). The abstract should contain a concise description of the econometric/economic problem you analyse and of your results. This allows the busy reader to obtain quickly a clear idea of the thesis content.

Contents

List of Abbreviations	i
List of Figures	ii
List of Tables	iii
1 Introduction	1
1.1 EGFR Signaling Cascade	4
1.2 EGFR Signaling in Cancer	5
1.3 Targeted Sequencing and Target Enrichment Methods	7
1.4 Sequencing Chemistries	8
1.4.1 Illumina MiSeq	8
1.5 NGS Data Analysis	11
1.6 Practical Implications in the Laboratory	11
1.7 Aims of the Thesis	15
2 Material Methods	15
2.1 Library Preparation	16
2.1.1 Patients	16
2.1.2 DNA Extraction, Quantification Quality Control	16
2.1.3 Agilent Haloplex ClearSeq Cancer	16
2.1.4 Illumina TruSight Tumor 15	16
2.2 Bioinformatic Analysis	16
2.2.1 Agilent SureCall	16
2.2.2 Illumina BaseSpace TruSight Tumor 15 App	16
2.2.3 Custom In-House Pipeline (Velona)	16

2.2.4	Variant Calling Algorithms	17
3	Results	18
3.1	Sample Preparation	18
3.2	NGS Data Quality	18
3.3	Coverage Analysis	20
3.4	Variant Calling Algorithm Comparison	21
4	Conclusions	22
	References	23

List of Abbreviations

NGS	Next Generation Sequencing	LNS	Laboratoire National de Sante
SGMB	Service of Genetics and Molecular Biology	TST15	Illumina TruSight Tumor 15

List of Figures

1	Schematic representation of the EGFR signaling cascade [1]	6
2	Electropherograms of representative sequencing libraries prepared by Agilent Halo-plex ClearSeq Cancer and Illumina TruSight Tumor 15. (*) represents the lower marker, (**) represents the upper marker	18
3	Scatter plot of the corrected peak area (X axis) of the regions corresponding to the sequencing libraries defined in the blabla software and the dCt (Y axis). Agilent Halo-plex ClearSeq Cancer data are represented as blue dots, Illumina TruSight Tumor 15 data are represented as red dots.	19
4	Comparison of coverage distributions per amplicon as reported by FastQC	20
5	Comparison of Coverage Distributions per Amplicon	21

List of Tables

1	ISV	19
2	Failed Amplicons in Agilent Haloplex ClearSeq Cancer	21
3	Failed Amplicons in Illumina TruSight Tumor 15	21

1 Introduction

- What is the subject of the study? Describe the economic/econometric problem.
 - ¿ Implementation of Cancer Gene Panels Variant Calling Algorithms
- What is the purpose of the study (working hypothesis)?
 - ¿ Check whether one of them is better than the others. -¿ Study effect of FFPE
- What do we already know about the subject (literature review)? Use citations: [2] shows that... Alternative Forms of the Wald test are considered [?].
- What is the innovation of the study?
 - ¿ relatively rare technique in diagnostics, but has many advantages
- Provide an overview of your results.

||||| HEAD ¿¿¿ Internship Report

The accumulation of DNA mutations is a characteristic of tumors. Mutations in a cell's genetic material enable the so-called cancer hallmarks, which include, amongst others, capabilities as activating invasion and metastasis, enabling replicative immortality, sustaining proliferative signaling or evading growth suppressors. Classical anti-cancer treatments are tailored for the average patient“ and not for the individual. Traditional cytotoxic chemotherapeutic drugs, for instance, are unspecific and have numerous adverse effects [1]: they attack rapidly dividing cells and make no difference between healthy and cancer cells. For a long time, there was no possibility to predict the success of a patient's cancer treatment. In consequence, the treating clinician had no way to personalize the treatment to the individual patient.

The completion of the human genome project and technical advances in the field of genetics, in combination with advanced computer performance, enabled geneticists and pharmacologists to adopt the personalization of cancer diagnostics and treatment. In oncology, the customized health-care approach is designed to the individual patient by taking into account the genetic information from a patient's tumor biopsy. Targeting proteins that give cancer cells a proliferative advantage, instead of simply using cytotoxic agents, allows a more specific treatment and might decrease treatment-associated side-effects. In the past years, the Food and Drug Administration (FDA, USA) has approved several drugs that specifically target proteins needed for cancerogenesis [2]. Many of these drugs are now competent treatments for several common cancers. In many patients, however,

these agents might lack any efficiency: cancers are known to be highly heterogeneous and mutations in some subpopulations of cancer cells make them resistant to the treatment [3]. Identifying these mutations is thereby an utmost necessity for tailoring a targeted and efficient cancer therapy. Amongst the many mutation detection techniques, Next Generation Sequencing (NGS) constitutes the most powerful method and allows deep insights into the underlying causes of diseases.

ۛۛۛ Next-Generation Sequencing: Methodology and Application Nucleic acid sequencing is a method for determining the exact order of nucleotides present in a given DNA or RNA molecule. In the past decade, the use of nucleic acid sequencing has increased exponentially as the ability to sequence has become accessible to research and clinical labs all over the world. The first major foray into DNA sequencing was the Human Genome Project, a

Since completion of the first human genome sequence, demand for cheaper and faster sequencing methods has increased greatly. This demand has driven the development of second-generation sequencing methods, or next-generation sequencing (NGS). NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced in unison. Massively parallel sequencing technology facilitates high-throughput sequencing, which allows an entire genome to be sequenced in less than one day. In the past decade, several NGS platforms have been developed that provide low-cost, high-throughput sequencing. Here we highlight two of the most commonly used platforms in research and clinical labs today: the Life Technologies Ion Torrent Personal Genome Machine (PGM) and the Illumina MiSeq. The creation of these and other NGS platforms has made sequencing accessible to more labs, rapidly increasing the amount of research and clinical diagnostics being performed with nucleic acid sequencing.

ۛۛۛ Clinical Actionability Enhanced through Deep Targeted Sequencing of Solid Tumors Next generation sequencing can facilitate personalized cancer therapy approaches by identifying actionable somatic events in tumor samples (1). Furthermore, high-quality sequencing data can reveal associations with sensitivity or resistance that can inform the development and implementation of targeted therapeutics and, in particular, aid in the design of future trials to validate findings and actionability. Critical alterations include mutations [single nucleotide variations (SNVs)¹⁴ and insertions and deletions (indels)], copy number variations (CNVs), and rearrangements that can potentially predict response and resistance to targeted agents. Whole genome sequencing (WGS) and whole exome sequencing (WES) allow the detection of SNVs, indels, CNVs, and rearrangements. However, the relatively low coverage of WGS and WES, as currently implemented in most of the sequencing laboratories (100–250), may have limited ability to cost-effectively detect aberrations that are present in a subpopulation of tumor cells while identifying a myriad of aberrations of unknown

clinical significance (2). Somatic aberrations present at low allele frequencies across different types of tumors (3, 4) can potentially impact patient prognosis or response (5) and thus are important to detect reliably. Targeted sequencing to a depth that allows detection of relatively low mutant allele frequency (MAF) may represent an alternative or a complement to WGS and WES to detect clinically relevant alterations. Additionally, in most clinical and research settings, the amount of DNA that can be isolated from tumor samples is limited and the DNA is often damaged owing to fixation and storage procedures such as those used with formalin-fixed paraffin-embedded (FFPE) samples. Therefore a multiplexed targeted platform that can generate reliable data with high sensitivity from limited amounts of DNA from FFPE samples is needed. Several targeted sequencing panels have been successfully implemented (6, 7). However, the details of a platform's design and parameterization will influence the precision and reliability of the molecular profiling results, impacting both translational research and clinical decision-making. Thus, it is of great value to explore multiple potential solutions in a real patient care environment until a community-wide solution is established, validated, and well accepted.

1: Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci Transl Med* 2011;3:111ra21. 5: Zhao BY, Pritchard JR, Lauffenburger DA, Hemann MT. Addressing genetic tumor heterogeneity through computationally predictive combination therapy. *Cancer Discov* 2014;4:166–74. 6: Wagle N, Berger MF, Davis MJ, Blumenstiel B, Defelice M, Pochanard P, et al. High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. *Cancer Discov* 2012;2:82–93. 7: Frampton GM, Fichtenholtz A, Otto GA, Wang K, Downing SR, He J, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* 2013;31:1023–31.

🔗🔗 Next-generation sequencing-based multi-gene mutation profiling of solid tumors using fine needle aspiration samples: promises and challenges for routine clinical diagnostics

Personalized therapy by identification and targeting of tumor-specific molecular abnormalities is rapidly becoming an important component in the management of cancer patients. Consequently, algorithms for tumor diagnosis necessitate not only morphological and immunophenotypic assessment of tumors but also molecular mutational profiling. Among the solid tumors, mutational status is important to the clinical management of patients with thyroid carcinomas, non-small cell carcinomas of the lung, melanomas and colorectal carcinomas. For example, treatment of patients with non-small cell carcinoma of the lung depends on the histological type (adenocarcinoma versus squamous cell carcinoma) and the mutational status of the epidermal growth factor receptor (EGFR) and

KRAS genes.¹ Thus, adequate tissue sampling is essential not only for pathological diagnosis but also for molecular testing that is required to guide therapeutic decisions.^{1, 2, 3} 2: Hirsch FR, Wynes MW, Gandara DR et al. The tissue is the issue: personalized medicine for non-small cell lung cancer. *Clin Cancer Res* 2010;16:4909–4911. 3: Cagle PT, Allen TC, Dacic S et al. Revolution in lung cancer: new challenges for the surgical pathologist. *Arch Pathol Lab Med* 2011;135:110–116.

In clinical practice, minimally invasive fine needle aspiration (FNA) is a helpful and convenient method for establishing the diagnosis of solid tumors. FNA procedures are included in the recommended guidelines for the diagnosis of thyroid carcinomas,^{4, 5} lung carcinomas^{6, 7} and sarcomas, such as Ewing Sarcoma/primitive neuroectodermal tumor.⁸ Although the DNA yield obtained by FNA is adequate for morphological diagnosis, FNA material is not routinely processed for molecular analysis.⁹ This is because mutational assessment of multiple genes by traditional sequencing platforms require a large quantity of DNA, which is often difficult to obtain by FNA. Fortunately, most FNA procedures routinely have either a concurrent biopsy or a follow-up surgical excision, which provide adequate tissue for molecular assessment.^{10, 11} 6: Aisner DL, Deshpande C, Baloch Z et al. Evaluation of EGFR mutation status in cytology specimens: an institutional experience. *Diagn Cytopathol* 2013;41:316–323. 7: Ulivi P, Romagnoli M, Chiadini E et al. Assessment of EGFR and K-ras mutations in fixed and fresh specimens from transesophageal ultrasound-guided fine needle aspiration in non-small cell lung cancer patients. *Int J Oncol* 2012;41:147–152. 10: Santis G, Angell R, Nickless G et al. Screening for EGFR and KRAS mutations in endobronchial ultrasound derived transbronchial needle aspirates in non-small cell lung cancer using COLD-PCR. *PloS One* 2011;6:e25191.

In some instances, however, cytological specimens may be the only material available for molecular testing. If the tumor is locally advanced or metastatic, surgical resection is not performed.^{1, 6} In some patients, tumor size, location or co-morbid conditions may preclude concurrent core needle or excisional biopsy.¹

1.1 EGFR Signaling Cascade

===== Cancer is a major public health problem worldwide and one of the leading death causes. In 2012, there were an estimated 14.1 million new cancer cases with estimated 8.2 million cancer deaths [3]. Lung cancer is the most common cancer, both in terms of new cases (1.8 million) and deaths (1.6 million). Breast cancer is the second most common cancer (1.7 million cases) but only ranks 5th as cause of death (522,000 deaths). Colorectal cancer (1.4 million cases; 694,000

deaths), prostate cancer (1.1 million cases; 307,000 deaths), stomach cancer (951,000 cases; 723,000 deaths) and liver cancer (782,000 cases; 723,000 deaths) are following.

In Luxembourg, there were 1164 death cases caused by cancer in 2014, accounting for 30.6% of the number of deaths. Cancer is thereby the second most common death cause in Luxembourg, ranging behind cardiovascular diseases (1189 deaths; 31.3%). Amongst these cancer deaths, cancers of the digestive organs (340 deaths) and of the respiratory organs (272 deaths) were the most frequent ones. [4].


(Considerable effort is made by the research community to understand the underlying causes of cancer and its progression, but death rate related to cancer remains high. In Luxembourg, the death rate did not significantly drop from 1998 to 2014.)

In recent years, more targeted drugs. personalized medicine.

Next-Generation Sequencing (NGS) is a powerful method to analyze RNA or DNA molecules. Improved protocols and methods have been developed in recent years that allow to implement NGS for a variety of applications in both research and diagnostics. NGS is still hard to implement and is mainly used in research. Only a few laboratories are using this technique as a diagnostic technique. The SGMB is building expertise in NGS and is implementing targeted NGS in its molecular pathology laboratory. This implementation aims at gaining deep insights into the underlying causes of solid tumors and in guiding targeted cancer therapy.

The goal of this master's thesis was to compare commercially available cancer gene panels on samples of solid tumor patients by several parameters and by their ability to detect variants, which have previously been detected in the routine mutation detection workflow in the laboratory. As a second part, several freely available variant calling algorithms were compared for their potential implementation in the custom in-house variant detection bioinformatic pipeline.

1.2 EGFR Signaling in Cancer

Bla [1]  a5147a619d3523c598b91a86e2c62e19a96fd26f

Describe pathway

Common mutations in this pathway

EGFR-targeted drugs: bevacizumab, cetuximab, panitumumab, ...

The samples analyzed in the SGMB originate from patients suffering from solid tumors, mainly

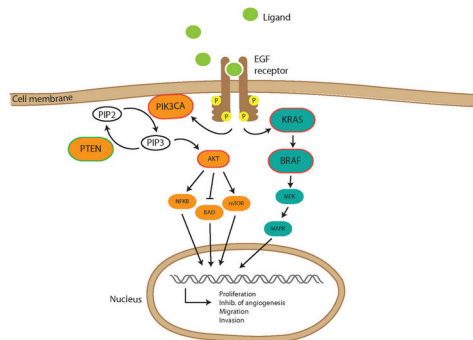


Figure 1: Schematic representation of the EGFR signaling cascade [1]

melanoma, non-small cell lung carcinoma (NSCLC), and colorectal cancer. Evidence suggests that many solid tumors use and modify EGFR (Epithelial Growth Factor Receptor) signaling for their purposes [4]. Targeting this signaling pathway is thereby an attractive anti-cancer treatment. In this regard, anti-EGFR monoclonal antibodies (cetuximab (Erbix[®]) and panitumumab (Vectibix[®]) and tyrosine kinase inhibitors (erlotinib (Tarceva[®]) and gefitinib (Iressa[®])) have shown their usefulness in cancer treatments [5]. EGFR and downstream proteins K-Ras / N- Ras and B-Raf are predictive biomarkers for the successfulness of the administration of the mentioned drugs [6]. Comprehensive information about these markers is thereby essential when choosing a suitable treatment in order to minimize treatment- associated side-effects and to maximize the benefits of the treatment.

In their article, Scaltriti and Baselga present the EGFR signaling pathway as a model for targeted therapy [7]. EGFR is part of the family of receptor tyrosine kinases. This transmembrane protein is composed of an intracytoplasmic tyrosine kinase domain, a short hydrophobic transmembrane region and an extracellular ligand-binding domain. Upon ligand (EGF, TGF) binding, EGFR becomes activated. This leads to homodimerization, which results in an auto- and cross-phosphorylation of key tyrosine residues on its cytoplasmic domain. This forms docking sites for cytoplasmic proteins that contain phosphotyrosine-binding and Src homology 2 domains. This allows, amongst others, signaling through the PTEN/PI3K/AKT and RAS-RAF-MAPK pathways. Activation of PTEN/PI3K/AKT leads to cell growth, proliferation and survival [8], while RAS-RAF-MAPK induces cell survival and cell cycle progression and proliferation [9]. In the RAS-Raf-MAPK pathway, Grb2 and Sos, two adaptor proteins, form a complex with the activated EGFR [10]. The resulting conformational change of Sos recruits Ras-GDP, which in turn becomes activated to form Ras-GTP. Ras-GTP activates Raf, which, in intermediate steps, phosphorylates a MAPK (mitogen-activated protein kinase). Activated MAPKs are then imported from the cytoplasm into the nucleus where they act on target genes. The Ras and Raf subfamilies include several proteins, three of them are of

interest in targeted cancer treatments: K-Ras, N-Ras and B-Raf.

Activating mutations of EGFR or its downstream proteins provide resistance to specific treatments. In that regard, the Institut National du Cancer (F) [11] provides recommendations about which mutations have to be searched to identify patients eligible for the administration of monoclonal antibodies or tyrosine kinase inhibitors. For instance, mutations on codon 12 and 13 in the KRAS gene provide resistance to the monoclonal antibody agents panitumumab and cetuximab. Gefinitinib, a tyrosine kinase inhibitor, can only be prescribed for patients, which show no activating mutations on EGFR.

1.3 Targeted Sequencing and Target Enrichment Methods

||||| HEAD Several NGS bench-top devices have become available in the last decade. These instrumentations differ in their underlying chemistry that influences the instrument's performance, accuracy, output and time per run. Common sequencing principles include pyrosequencing (454), sequencing by ligation (SOLiD), ion semiconductor (Ion Torrent) and sequencing by synthesis (Illumina) [12]. Even though advances in sequencing technology and computational power and tools have decreased the time and cost of a sequencing experiment, NGS is still mainly used in research, with only a few laboratories using this technique in diagnostics.

In fact, validation of a NGS methodology requires careful assessment of methods and tools [13]. Therefore, each step that is performed from the initial starting material to sample processing, sequencing library preparation, sequencing assay and bioinformatic processing has to be carefully checked for sources of potential errors or variability. Basically, in the validation process, it is checked whether the method measures what it claims to measure with the required sensitivity and sensibility.

With the success of NGS, many cancer genomes have been sequenced and made available to the worldwide research community. Companies, molecular diagnostics laboratories and academic centers are trying to use these big data for their purposes. A lot of mutations are described in these genomes, but only a small fraction of them are clinically actionable, e.g. can be targeted with specific drugs. Therefore, a molecular pathology laboratory does not need to perform whole-genome or -exome sequencing, but can employ targeted NGS to analyze some genes of interest, which include mutations for which there exists a clinical utility. Due to the low number of target regions, targeted NGS allows high coverage. In addition, it is a time- and cost- efficient alternative to whole-genome or -exome sequencing. Also, targeted NGS results in a significantly lower amount of produced data and thereby eases data storage and analysis time. Table 1 shows a selection of

commercially available cancer gene panels, which all allow to analyze selected regions of genes implicated in cancerogenesis. Before implementing one of these panels for molecular diagnostics, it has to be ensured that the panel allows to study the genes of interest, e.g. genes that are clinically applicable, and a careful assessment of its analytical validity has to be performed. ===== Sanger Sequencing

Human Genome Project

Benchtop sequencing devices: more labs can now do sequencing as it becomes faster cheaper more reliable.

More more databases describing SNPs known pathogenic hotspots

targeted sequencing as um kommen, cost-effective fast high-coverage

Bla [5]          a5147a619d3523c598b91a86e2c62e19a96fd26f

Hybrid capture

Selective circularization

PCR amplification

1.4 Sequencing Chemistries

IonTorrent, etc

1.4.1 Illumina MiSeq

<http://www.illumina.com/documents/products/techspotlights/techspotlightsequencing.pdf>

Illumina sequencing technology leverages clonal array formation and proprietary reversible terminator technology for rapid and accurate large-scale sequencing. The innovative and exible sequencing system enables a broad array of applications in genomics, transcriptomics, and epigenomics.

Sequencing templates are immobilized on a proprietary ow cell surface (Figure 1) designed to present the DNA in a manner that facilitates access to enzymes while ensuring high stability of surface- bound template and low non-specific binding of uorescently labeled nucleotides. Solid-phase amplification (Figures 2–7) creates up to 1,000 identical copies of each single template

molecule in close proximity (diameter of one micron or less). Because this process does not involve photolithography, mechanical spotting, or positioning of beads into wells, densities on the order of ten million single-molecule clusters per square centimeter are achieved.

The Illumina sequencing approach is built around a massive quantity of sequence reads in parallel. Deep sampling and uniform coverage is used to generate a consensus and ensure high confidence in determination of genetic differences. Deep sampling allows the use of weighted majority voting and statistical analysis, similar to conventional methods, to identify homozygotes and heterozygotes and to distinguish sequencing errors. Each raw read base has an assigned quality score so that the software can apply a weighting factor in calling differences and generating confidence scores.

Illumina data collection software enables users to align sequences to a reference in resequencing applications (Figure 13). Developed in collaboration with leading researchers, this software suite includes the full range of data collection, processing, and analysis modules to streamline collection and analysis of data with minimal user intervention. The open format of the software allows easy access to data at various stages of processing and analysis using simple application program interfaces.

https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Illumina dye sequencing is a technique used to determine the series of base pairs in DNA, also known as DNA sequencing. The reversible terminated chemistry concept was invented by Bruno Canard and Simon Sarfati at the Pasteur Institute in Paris.[1][2] It was developed by Shankar Balasubramanian and David Klenerman of Cambridge University,[3] who subsequently founded Solexa, a company later acquired by Illumina. This sequencing method is based on reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands. It can also be used for whole-genome and region sequencing, transcriptome analysis, metagenomics, small RNA discovery, methylation profiling, and genome-wide protein-nucleic acid interaction analysis.[4][5]

(among net :) The first step after DNA purification is fragmentation. Enzymes called transposomes randomly cut the DNA into short segments ("tags"). Adapters are added on either side of the cut points (ligation). Strands that fail to have adapters ligated are washed away.[6])

The next step is called reduced cycle amplification. During this step, sequences for primer binding, indices, and terminal sequences are added. Indices are usually six base pairs long and are used during DNA sequence analysis to identify samples. Indices allow for up to 96 different samples to be run together. During analysis, the computer will group all reads with the same index together.[7][8]

The terminal sequences are used for attaching the DNA strand to the flow cell. Illumina uses a “sequence by synthesis” approach.[8] This process takes place inside of an acrylamide-coated glass flow cell.[9] The flow cell has oligonucleotides (short nucleotide sequences) coating the bottom of the cell, and they serve to hold the DNA strands in place during sequencing. The oligos match the two kinds of terminal sequences added to the DNA during reduced cycle amplification. As the DNA enters the flow cell, one of the adapters attaches to a complementary oligo.

Once attached, cluster generation can begin. The goal is to create hundreds of identical strands of DNA. Some will be the forward strand; the rest, the reverse. Clusters are generated through bridge amplification. Polymerases move along a strand of DNA, creating its complementary strand. The original strand is washed away, leaving only the reverse strand. At the top of the reverse strand there is an adapter sequence. The DNA strand bends and attaches to the oligo that is complementary to the top adapter sequence. Polymerases attach to the reverse strand, and its complementary strand (which is identical to the original) is made. The now double stranded DNA is denatured so that each strand can separately attach to an oligonucleotide sequence anchored to the flow cell. One will be the reverse strand; the other, the forward. This process is called bridge amplification, and it happens for thousands of clusters all over the flow cell at once.

Over and over again, DNA strands will bend and attach to oligos. Polymerases will synthesize a new strand to create a double stranded segment, and that will be denatured so that all of the DNA strands in one area are from a single source (clonal amplification). Clonal amplification is important for quality control purposes. If a strand is found to have an odd sequence, then scientists can check the reverse strand to make sure that it has the complement of the same oddity. The forward and reverse strands act as checks to guard against artifacts. Because Illumina sequencing uses polymerases, base substitution errors have been observed,[10] especially at the 3' end.[11] Paired end reads combined with cluster generation can confirm an error took place. The reverse and forward strands should be complementary to each other, all reverse reads should match each other, and all forward reads should match each other. If a read is not similar enough to its counterparts (with which it should be a clone), an error may have occurred. A minimum threshold of 97

At the end of bridge amplification, all of the reverse strands are washed off the flow cell, leaving only forward strands. Primers attach to the forward strands and add fluorescently tagged nucleotides to the DNA strand. Only one base is added per round. A reversible terminator is on every nucleotide to prevent multiple additions in one round. Each of the four bases has a unique emission, and after each round, the machine records which base was added. This process is “sequence by synthesis.”

Once the DNA strand has been read, the strand that was just added is washed away. Then, the

index 1 primer attaches, polymerizes the index 1 sequence, and is washed away. The strand forms a bridge again, and the 3' end of the DNA strand attaches to an oligo on the flow cell. The index 2 primer attaches, polymerizes the sequence, and is washed away.

A polymerase sequences the complementary strand on top of the arched strand. They separate, and the 3' end of each strand is blocked. The forward strand is washed away, and the process of sequence by synthesis repeats for the reverse strand.

The sequencing occurs for millions of clusters at once, and each cluster has 1,000 identical copies of a DNA insert.[10] The sequence data is analyzed by finding fragments with overlapping areas, called contigs, and lining them up. If a reference sequence is known, the contigs are then compared to it for variant identification.

This piecemeal process allows scientists to see the complete sequence even though an unfragmented sequence was never run; however, because Illumina read lengths are not very long [11] (HiSeq sequencing can produce read lengths around 90 bp long [7]), it can be a struggle to resolve short tandem repeat areas.[7][10] Also, if the sequence is de novo and so a reference doesn't exist, repeated areas can cause a lot of difficulty in sequence assembly.[10] Additional difficulties include base substitutions (especially at the 3' end of reads[11]) by inaccurate polymerases, chimeric sequences, and PCR-bias, all of which can contribute to generating an incorrect sequence.[11]

1.5 NGS Data Analysis

GATK best practices

1.6 Practical Implications in the Laboratory

FFPE : more details

||||| HEAD The quality of the genetic testing of the tumor is affected by several factors. These include the content of tumor cells in the sample, the quality of the tissue material, sequencing library preparation and the the bioinformatic pipeline.

The biopsy usually consists of healthy and cancer cells. The sensitivity of tumor variant detection is linked to the tumor cell content of the specimen. In addition, cancers are highly heterogenous, e.g. a small subpopulation might present mutations that provide resistance to targeted treatment. Detecting these low-frequency mutations and clearly separating them from eventual high-frequency

fixation or sequencing artifacts presents a huge challenge [14].

Tumor biopsies usually yield a limited amount of tissue, therefore it is conceivable to use the same sample for more analyses. In Luxembourg, all relevant tumor biopsies are usually sent to the Laboratoire National de Sante (LNS) to the Service of Pathologic Anatomy where the biopsy is fixed in formalin and embedded in paraffin (FFPE). FFPE conserves the tissue morphology and thereby allows histological analysis. In addition, it allows to store specimens for decades. Sample quality, however, is influenced by this fixation method, but also by the size of the biopsy, and its fixation time [13]. DNA extraction from FFPE samples is difficult and yields low amounts of DNA [15]; formaldehyde leads to cross-linking of nucleic acids and proteins [16]; FFPE introduces fixation artifacts into DNA sequences [17—]. These circumstances complicate sample processing as well as NGS data interpretation. Though, FFPE samples have been shown to be still suitable for downstream analyses [18].

Sequencing library preparation also affects the final NGS result. Several technologies for target enrichment exist and are available for different sequencing instruments [19]. Essential for all these enrichment methods is the amplification of target regions and the introduction of multiplexing, which requires the incorporation of a unique index adaptor combination for each sample. Target enrichment methods can be separated into three basic groups: targeted circularization, hybrid capture of target fragments and PCR-based enrichment methods. PCR-driven methods happen on high-molecular DNA. In contrast to uniplex long-range PCR, short-range multiplex PCR produces short DNA fragments of target regions. There is thereby no need of DNA shearing. Hybridization-based methods require a so-called shotgun library construction before target regions can be captured. During this process, genomic DNA is sheared randomly into small fragments and an adapter- and index-linked library is produced. Biotinylated baits are added that bind to target regions. Target regions can then be captured using streptavidin coated magnetic beads. Targeted circularization methods rely on a digestion of DNA by restriction enzymes. The produced DNA fragments are then circularized and uncircularized DNA fragments are removed by exonucleases. Only circularized target regions are then amplified by PCR.

The establishment and validation of a bioinformatic NGS data analysis pipeline still constitutes a challenge in diagnostics. After generation of FASTQ files of the sequencer, data generally undergo quality control, followed by trimming of low quality bases, alignment to the reference genome, variant calling and variant annotation. For each of these steps, several bioinformatic algorithms and tools exist [20]. The computational pipeline of the molecular pathology laboratory has to incorporate the tools that allow the most sensitive and sensible analysis of data. For instance, quality trimming

influences the mapping to the reference genome. The mapping, in turn, strongly affects the variant calling. In fact, variant calling is a critical step in NGS data analysis. Several tool kits as SAMtools, SPLINTER, VarScan2 or GATK allow variant annotation, but vary in their false-positive and false-negative detection rates ([21], [22]). These tools have to be carefully assessed, as false-positives or false-negatives should absolutely be avoided when it comes to the subscription of a targeted chemotherapeutic agent.

To facilitate interpretation of NGS data, variants have to be annotated and their clinical actionability has to be identified. Several databases have emerged in this field (such as mycancergenome.org) and numerous tools allow to automatize variant annotation. Here again, the choice of the database and the variant annotator is important.

Finally, the sample-to-results time is a very pragmatic, but important factor. The time from the biopsy to the potential start of an administration of a targeted chemotherapeutic drug should be reduced to a minimum. For instance, in case of late-stage cancer patients, it would be unacceptable if analysis would take several weeks. To reduce the sample-to-results time to under two weeks, the sample processing workflow should be as short as possible, while still yielding high quality sequencing libraries. The bioinformatic pipeline should not only incorporate the best tools, but should also be automatized to further reduce the time of analysis. ===== Chemical fixatives are used to preserve tissue from degradation, and to maintain the structure of the cell and of sub-cellular components such as cell organelles (e.g., nucleus, endoplasmic reticulum, mitochondria). The most common fixative for light microscopy is 10 phosphate buffered saline).

These fixatives preserve tissues or cells mainly by irreversibly cross-linking proteins. The main action of these aldehyde fixatives is to cross-link amino groups in proteins through the formation of methylene bridges (-CH₂-), in the case of formaldehyde.

This process, while preserving the structural integrity of the cells and tissue can damage the biological functionality of proteins, particularly enzymes, and can also denature them to a certain extent. This can be detrimental to certain histological techniques.

Formalin fixation leads to degradation of mRNA, miRNA and DNA in tissues. However, extraction, amplification and analysis of these nucleic acids from formalin-fixed, paraffin-embedded tissues is possible using appropriate protocols.

The aim of Tissue Processing is to remove water from tissues and replace with a medium that solidifies to allow thin sections to be cut. Biological tissue must be supported in a hard matrix to allow sufficiently thin sections to be cut, typically 5 μ m (micrometres; 1000 micrometres = 1 mm) thick

for light microscopy.

Since it is immiscible with water, the main constituent of biological tissue, water must first be removed in the process of dehydration. Samples are transferred through baths of progressively more concentrated ethanol to remove the water. This is followed by a hydrophobic clearing agent (such as xylene) to remove the alcohol, and finally molten paraffin wax, the infiltration agent, which replaces the xylene. Paraffin wax does not provide a sufficiently hard matrix for cutting very thin sections for electron microscopy. Instead, resins are used. Epoxy resins are the most commonly employed embedding media, but acrylic resins are also used, particularly where immunohistochemistry is required. Thicker sections (0.35 μ m to 5 μ m) of resin-embedded tissue can also be cut for light microscopy. Again, the immiscibility of most epoxy and acrylic resins with water necessitates the use of dehydration, usually with ethanol.

After the tissues have been dehydrated, cleared, and infiltrated with the embedding material, they are ready for external embedding. During this process the tissue samples are placed into molds along with liquid embedding material (such as agar, gelatine, or wax) which is then hardened. This is achieved by cooling in the case of paraffin wax and heating (curing) in the case of the epoxy resins. The acrylic resins are polymerised by heat, ultraviolet light, or chemical catalysts. The hardened blocks containing the tissue samples are then ready to be sectioned.

Because Formalin-fixed, paraffin-embedded (FFPE) tissues may be stored indefinitely at room temperature, and nucleic acids (both DNA and RNA) may be recovered from them decades after fixation, FFPE tissues are an important resource for historical studies in medicine.

For light microscopy, a steel knife mounted in a microtome is used to cut 4-micrometer-thick tissue sections which are mounted on a glass microscope slide. For transmission electron microscopy, a diamond knife mounted in an ultramicrotome is used to cut 50-nanometer-thick tissue sections which are mounted on a 3-millimeter-diameter copper grid. Then the mounted sections are treated with the appropriate stain.

quantitation

ffpe quality control necessary. FFPE are degraded. More C₂T

Target enrichment methods hun och vlt en effekt

choice vum sequencer, different error rates, different yield, analysis time.

choice vun den bioinformatic tools. t'gin der vill, vir- an nodeeler. zB variant calling algorithms: most are for haplotypes and not for somatic variants. many need matched tumor-normal samples,

which we here in the lab do not have, only have tumor dna. after variant calling, they have to be annotated for their clinical relevance and potential actionability.

problem: many variants may be found, some of them may be harmless germline SNPs, many variants may be causative of cancer, but are not actionable.

time sample-to-result

how to report? all results or only those that are known to be actionable. 5147a619d3523c598b91a86e2c

1.7 Aims of the Thesis

Targeted NGS is still not widely used in diagnostics laboratories. The SGMB of the LNS is planning to build expertise with the aim to adopt NGS routinely in the laboratory, mainly in the context of diagnosis and therapy of cancer patients in Luxembourg.

The aim of this thesis project was to test commercially available cancer gene panels, e.g. Illumina Trusight Tumor 15 and Agilent Haloplex HS ClearSeq Cancer, for their potential use in the routine workflow of the laboratory. Several samples of cancer patients were prepared with both kits and were sequenced on the Illumina MiSeq device. Both kits vary in their sequencing library preparation principles: Illumina's tst15 uses the multiplex PCR approach while Agilent's Haloplex Enrichment System uses enzymatic DNA restriction followed by probe capture.

NGS data were analyzed with the respective recommended pipelines and a custom in-house pipeline.

Finally, several freely available variant calling algorithms were tested for their potential implementation in the custom in-house variant discovery bioinformatic pipeline.

2 Material Methods

- How was the data analyzed ?
- Present econometric/statistical estimation method and give reasons why it is suitable to answer the given problem.
- Allows the reader to judge the validity of the study and its findings.
- Depending on the topic this section can also be split up into separate sections.

2.1 Library Preparation

2.1.1 Patients

Melanoma Non-Small Cell Lung Carcinoma (NSCLC) metastatic colorectal cancer (mCRC) Chronic lymphocytic leukemia (CLL) were extracted from blood and did not undergo FFPE treatment - they were used as some kind of good quality samples to see if there are really more C₂T variants in FFPE samples.

2.1.2 DNA Extraction, Quantification Quality Control

DNA Extraction Kit from Qiagen

Quantification Qubit fluorometer, either High Sensitivity kit or Broad Range

Quality Control Illumina Infinium FFPE QC Assay kit

2.1.3 Agilent Haloplex ClearSeq Cancer

2.1.4 Illumina TruSight Tumor 15

2.2 Bioinformatic Analysis

2.2.1 Agilent SureCall

With alignment algorithms installed Windows System, 3.00GHz, 16GB RAM\$ modifiable

2.2.2 Illumina BaseSpace TruSight Tumor 15 App

Cloud-based parameters not modifiable

2.2.3 Custom In-House Pipeline (Velona)

Linux System

2.2.4 Variant Calling Algorithms

Tested on Linux Ubuntu 14.04.4 LTS Trusty Tahr installed on VMware virtual box with of 12GB RAM

GATK HaplotypeCaller

VarScan 2

Mutect1.1.7 [2]

SomVarIUS

3 Results

3.1 Sample Preparation

Before pooling the adaptor-ligated and indexed sequencing libraries, the success of library preparation is validated using the Agilent Bioanalyzer instrument. 2a and 2b show representative electropherograms of a sample that has been processed using both kits. The expected DNA products should be detected at 175-600 bp for Haloplex CSC and 200-400 for TST15.

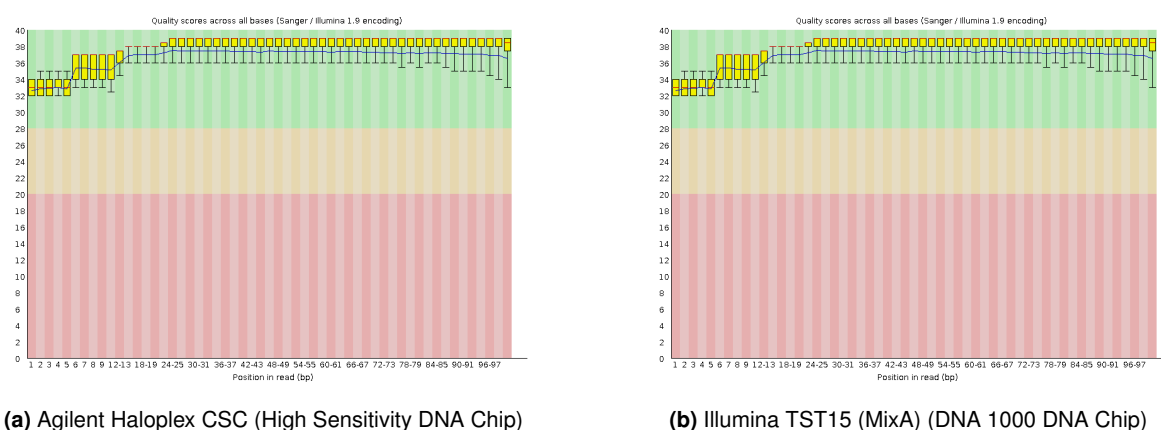


Figure 2: Electropherograms of representative sequencing libraries prepared by Agilent Haloplex ClearSeq Cancer and Illumina TruSight Tumor 15. (*) represents the lower marker, (**) represents the upper marker

Using the blablabla software, the concentration, molarity and total peak area (TPA) of the expected sequencing libraries were calculated.

Hopefully: there is a relationship between dCt and TPA. Maybe: Haloplex is more affected by bad dCt values.

3.2 NGS Data Quality

Sequencing run parameters were calculated by the Illumina Sequencing Viewer software. Table XXX shows the averaged run parameters of runs with Haloplex CSC and TST15 sample preparation.

TST15 has a higher cluster density and therefore a higher total yield, but has lower reads passing a phred-score threshold of Q30 than Haloplex. This is due to the different chemistries used. TST15 uses v3 chemistry, while Haloplex uses v2. v2 generally has lower cluster density and output, but therefore better quality.

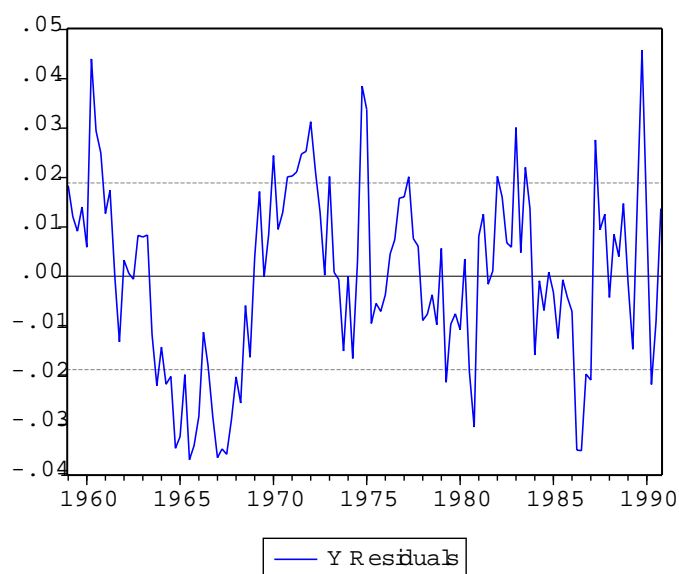


Figure 3: Scatter plot of the corrected peak area (X axis) of the regions corresponding to the sequencing libraries defined in the blabla software and the dCt (Y axis). Agilent Haloplex ClearSeq Cancer data are represented as blue dots, Illumina TruSight Tumor 15 data are represented as red dots.

Table 1: Comparison of Run Parameters (Averaged) of Sequencing Runs with Haloplex CSC & TST15 Sample Preparation

Parameter	Haloplex CSC	TST15
Yield total (Gb)	3.7	7.37
% >Q30	93.8	82.355
Cluster Density PF (k/mm2)	1084	1180
Cluster Density PF (%)	85.95	79.95

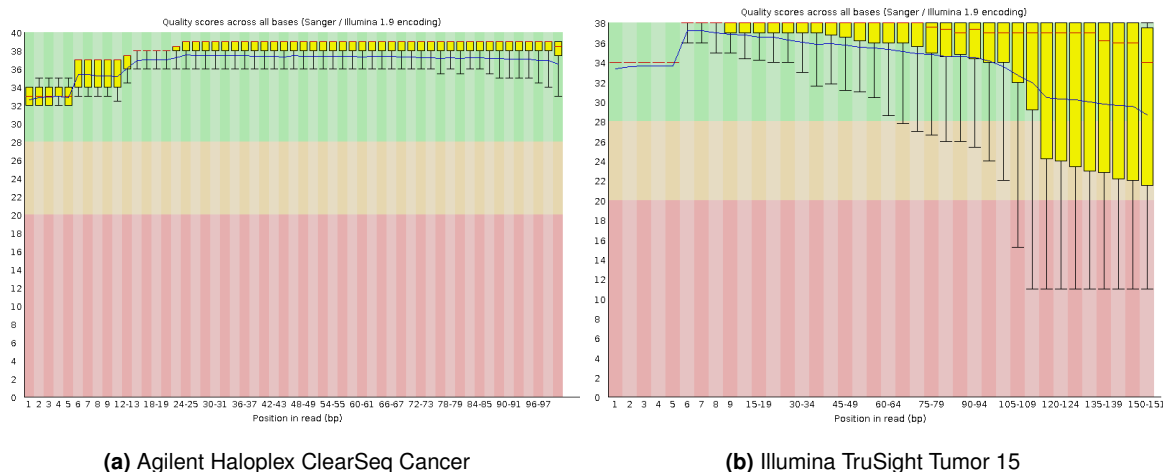


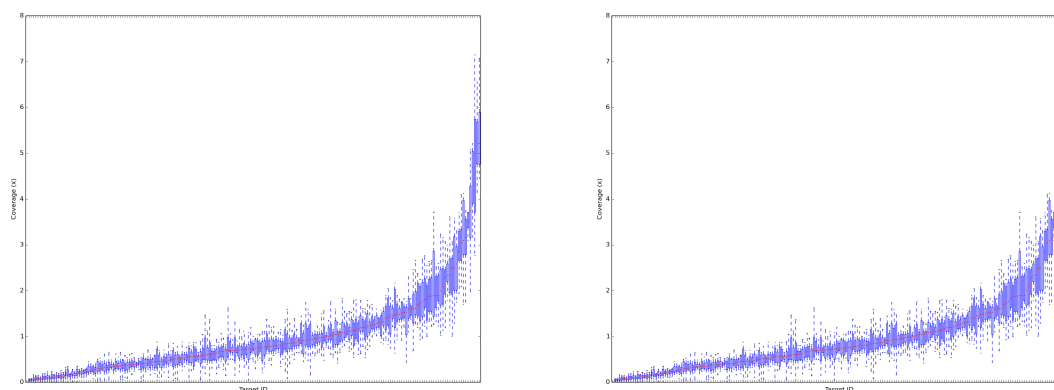
Figure 4: Comparison of coverage distributions per amplicon as reported by FastQC

Parameter	Haloplex TST15	
	CSC	
% mapped	91	62.6
% paired	0	58.7
% singletons	1.8	3.8

Table XXX shows the boxplot representations of the read qualities per position of two representative FASTQ files as reported by FastQC. Both workflows yield high quality data, yet Haloplex CSC data have more narrow distributions and are of higher quality. This is in direct relationship with the sequencing chemistry kit used.

The Samtools Flagstat command was used to determine some basic BAM statistics of BAM files of samples prepared with the respective library preparations and processed with the mentioned bioinformatic pipelines. ?? shows the averaged result of these statistics.

Considering the recommended pipelines, Haloplex CSC data, analyzed with Agilent's SureCall software, has a higher percentage (91%) of mapped reads when compared to Illumina's BaseSpace TruSight Tumor 15 App (62.9%). Data analysis with the recommended SureCall design includes a steps where mates are fixed, but they are not stitched together. Therefore no reads are considered as being paired. The TST15 app in contrast includes a read stitching step and 58% are considered as properly paired. This means that of the 62.6% of mapped reads, 4.2% are not properly paired. 3.8% of reads processed with the TST15 online App are considered to be singletons, whereas only



(a) Agilent Haloplex ClearSeq Cancer

(b) Illumina TruSight Tumor 15

Figure 5: Comparison of Coverage Distributions per Amplicon

1.8% of reads processed with the SureCall software are considered as singletons.

3.3 Coverage Analysis

Coverage Distribution TST15 vs Haloplex

Coverage Distribution per Patient (check if correlation IQR with dCt)

Coverage Distribution per Amplicon (check if some have always lower coverage, check if some failed)

Failed Amplicon Counter

Table 2: Failed Amplicons in Agilent Haloplex ClearSeq Cancer **Table 3:** Failed Amplicons in Illumina TruSight Tumor 15

Amplicon	Coverage Failed	No of Samples
ATM ₁ 4	1	all
Bla	200	9
Blabla	500	5

Amplicon	Coverage Failed	No of Samples
ATM ₁ 4	1	all
Bla	200	9
Blabla	500	5

Fragmentation $j-\hat{c}$ Coverage?

(GATK CallableLoci) (GATK CountLoci???) (GATK FindCoveredIntervals)

On-off target; Enrichment Efficiency TST15 vs Haloplex

Coverage across genome, check where there is coverage

Strandedness?

GATK DepthOfCoverage???

3.4 Variant Calling Algorithm Comparison

Detection of Known Single Nucleotide Variants and Deletions

Which should be found?

Which have not been found? Why?

MuTect

VarScan

GATK HaplotypeCaller

SomVarIUS???????? Freebayes???????? Vardict????????

SureCall TST15

Tools that may be of use somehow: GATK SelectVariants; GATK VariantFiltration; GATK VariantEval; GATK ValidateVariants

More C₂T ?

4 Conclusions

- Give a short summary of what has been done and what has been found.
- Expose results concisely.
- Draw conclusions about the problem studied. What are the implications of your findings?
- Point out some limitations of study (assist reader in judging validity of findings).
- Suggest issues for future research.

References

- [1] M. Berg, “EGFR and downstream genetic alterations in KRAS/BRAF and PI3K/AKT pathways in colorectal cancer – implications for targeted therapy,” *Discovery Medicine*, vol. 14, no. 76, pp. 207–214, 2012.
- [2] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz, “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples,” *Computational Biology*, 2013.
- [3] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, “Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012,” *International Journal of Cancer*, 2015.
- [4] R. d. d. Direction de la Santé, Service des statistiques, “Statistiques des causes de décès pour l’année 2014,” 2014.
- [5] F. Mertes, A. E. Sharawy, S. Sauer, J. M. L. M. van Helvoort, P. van der Zaag, A. Franke, M. Nilsson, H. Lehrach, and A. J. Brookes, “Targeted enrichment of genomic dna regions for next-generation sequencing,” *Briefings in Functional Genomics*, vol. 10, no. 6, pp. 374–386, 2011.