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

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

List of Figures

List of Tables

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: *Cibulskis et al. (2013a) shows that... Alternative Forms of the Wald test are considered (Breusch and Schmidt, 1988).*

- What is the innovation of the study?

  -¿ relatively rare technique in diagnostics, but has many advantages

- Provide an overview of your results.

General Intro

1.1 EGFR Signaling Cascade

The samples used in this study originate from samples that have been analyzed in the routine workflow of the SGMB. These samples originate from patients suffering from solid tumors, e.g. mainly melanoma, colorectal cancer (CRC) and non-small cell lung carcinoma (NSCLC).

There is evidence that the EGFR (Epithelial Growth Factor Receptor) signaling cascade is modified in those cancers.

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 (Erbitux) 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.

1.2 Targeted Sequencing and Target Enrichment Methods

Several NGS bench-top devices have become available in the last decade. These instrumentations differ in their underlying chemistry that influences the instruments 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.

1.3 Illumina MiSeq Sequencing Chemistry

1.4 NGS Data Analysis

GATK best practices

1.5 Practical Implications in the Laboratory

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 streptadivin 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.

1.6 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, with each having its specific benefits.

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

2.1.2 DNA Extraction, Quantification Quality Control

DNA Extraction

Quantification

Quality Control

2.1.3 Agilent Haloplex ClearSeq Cancer

2.1.4 Illumina TruSight Tumor 15

2.2 Bioinformatic Analysis

2.2.1 Agilent SureCall

Computer bla, 3.00GHz, 16GB RAM

2.2.2 Illumina BaseSpace TruSight Tumor 15 App

Cloud-based

2.2.3 Custom Pipeline (Velona)

Linux system

2.2.4 Variant Calling Algorithms

Mutect1.1.7 *Cibulskis et al. (2013b)*

SomVarIUS

VarScan 2

GATK HaplotypeCaller

3 Results

- Organize material and present results.
- Use tables, figures (but prefer visual presentation):
  - Tables and figures should supplement (and not duplicate) the text.
  - Tables and figures should be provided with legends.

    Figure 1 shows how to include and reference graphics. The graphic must be labelled before. Files must be in `.eps` *format.*

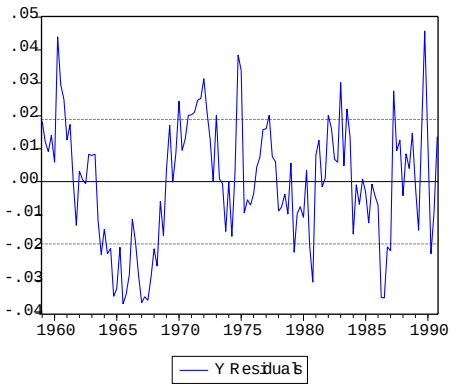


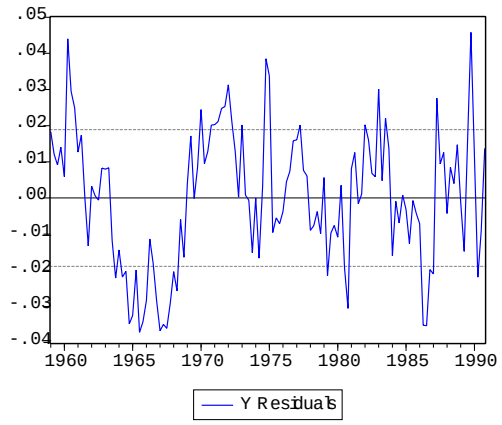
Figure 1: Estimated residuals from model XXX. ...

  - Tables and graphics may appear in the text or in the appendix, especially if there are many simulation results tabulated, but is also depends on the study and number of tables resp. figures. The key graphs and tables must appear in the text!
- Latex is really good at rendering formulas:

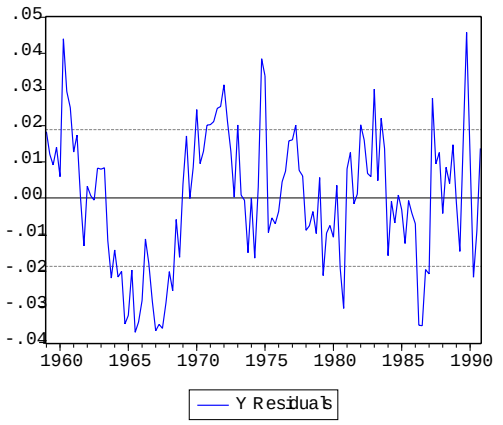
  Equation (1) represents the ACs of a stationary stochastic process:

$$f_y(\lambda) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\lambda j} = (2\pi)^{-1} \left(\gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j \cos(\lambda j) \right) \tag{1}$$

  where $i = \sqrt{-1}$ *is the imaginary unit,* $\lambda \in [-\pi, \pi]$ *is the frequency and the* γ_j *are the autocovariances of* y_t*.*



(a) Overlayed Electropherograms of Four Representative Libraries Prepared with Agilent Haloplex CSC



(b) Overlayed Electropherograms of Four Representative Libraries Prepared with Illumina TST15 (MixA & MixB)

Figure 2: Comparison of Agilent Bioanalyzer Electropherograms

- Discuss results:
  - Do the results support or do they contradict economic theory ?
  - What does the reader learn from the results?
  - Try to give an intuition for your results.
  - Provide robustness checks.
  - Compare to previous research.

3.1 Sample Preparation

FFPE ¡-¿ library concentration

Pictures of Bioanalyzer

3.2 NGS Data Quality

Illumina Sequencing Viewer

FASTQC

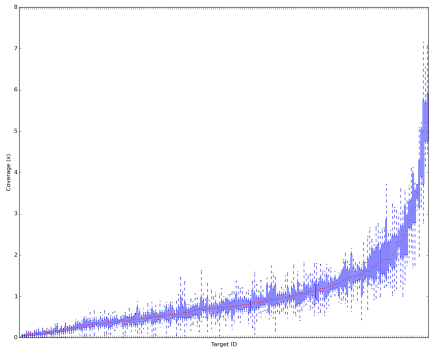
Table 1: Comparison of Run Parameters (Averaged) of Sequencing Runs with Haloplex CSC & TST15 Sample Preparation

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 (		

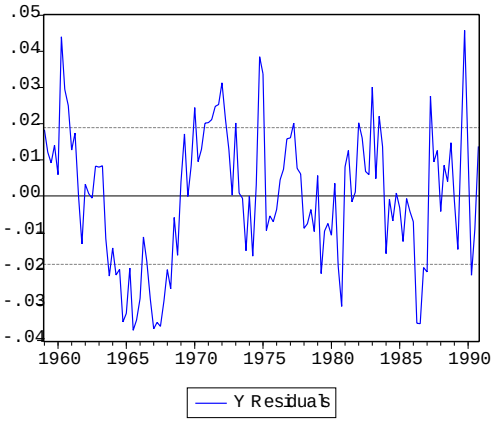
The Samtools Flagstat command was used to determine some basic BAM statistics of BAM files of samples prepared with the respective library preparations and processed with the mentioned bioinformatic pipelines. **??** shows the averaged result of these statistics. Considering the recommended pipelines, 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 1.8% of reads processed with the SureCall software are considered as singletons.

Table 2: Your caption here

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



(a) Agilent Haloplex ClearSeq Cancer



(b) Illumina TruSight Tumor 15

Figure 3: Comparison of Coverage Distributions per Amplicon

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 3: Failed Amplicons in Agilent Haloplex ClearSeq Cancer

Amplicon	Coverage Failed	No of Samples
ATM$_14$	1	all
Bla	200	9
Blabla	500	5

Table 4: Your caption here

Amplicon	Coverage Failed	No of Samples
ATM$_14$	1	all
Bla	200	9
Blabla	500	5

Fragmentation ¡-¿ 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?

Table 5: Known Variant Detection

Gene Ref TST15	Chr Alt	Pos Haloplex
Yield total (Gb)	3.7	7.37
% >Q30	93.8	82.355
Cluster Density PF (k/mm2)	1084	1180
Cluster Density PF (		

MuTect

VarScan

GATK HaplotypeCaller

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

SureCall TST15

GATK SelectVariants GATK VariantFiltration GATK VariantEval GATK ValidateVariants

More C¿T ?

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

BREUSCH, T. S. AND P. SCHMIDT (1988): "Alternative Forms of the Wald test: How Long is a Piece of String," *Communications in Statistics, Theory and Methods*, 17, 2789–2795.

CIBULSKIS, K., M. S. LAWRENCE, S. L. CARTER, A. SIVACHENKO, D. JAFFE, C. SOUGNEZ, S. GABRIEL, M. MEYERSON, E. S. LANDER, AND G. GETZ (2013a): "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Computational Biology*.

——— (2013b): "Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples," *Computational Biology*.