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: [1] shows that... Alternative Forms of the Wald test are considered [2].

- What is the innovation of the study?

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

- Provide an overview of your results.

General Intro; maybe subsection about

1.1 EGFR Signaling Cascade

Describe pathway

Common mutations in this pathway

EGFR-targeted drugs

1.2 Targeted Sequencing and Target Enrichment Methods

Hybrid capture

Selective circularization

PCR amplification

1.3 Illumina MiSeq Sequencing Chemistry

Picture

1.4 NGS Data Analysis

GATK best practices

1.5 Practical Implications in the Laboratory

FFPE : more details

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: 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

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 (author?) *[3]*

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.*

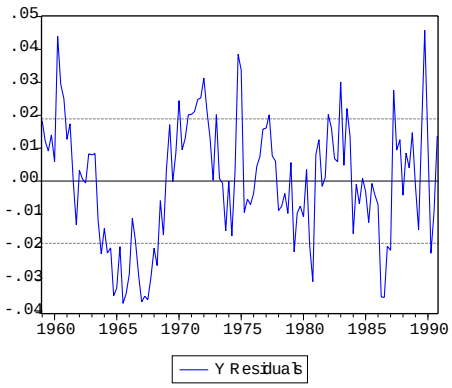


Figure 1: Estimated residuals from model XXX. ...

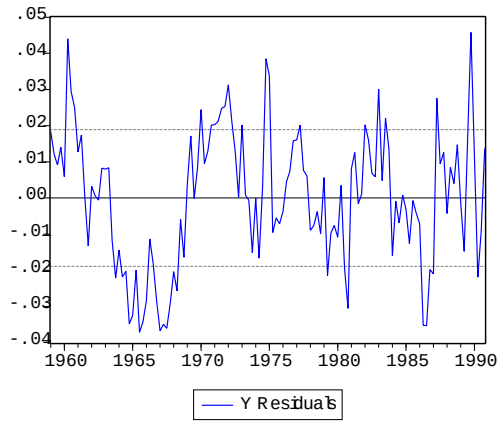
- Tables and graphics may appear in the text or in the appendix, especially if there are many simulation results tabulated, but is also depends on the study and number of tables resp. figures. The key graphs and tables must appear in the text!

- Latex is really good at rendering formulas:

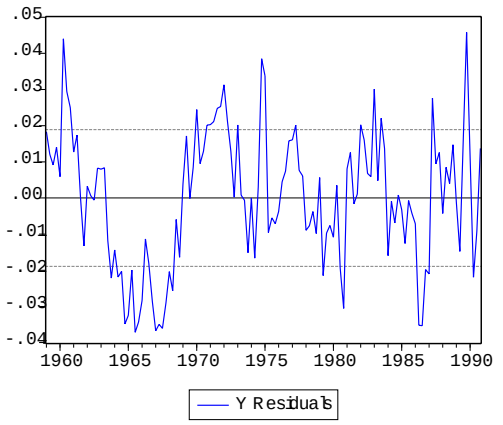
  Equation (1) represents the ACs of a stationary stochastic process:

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

  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

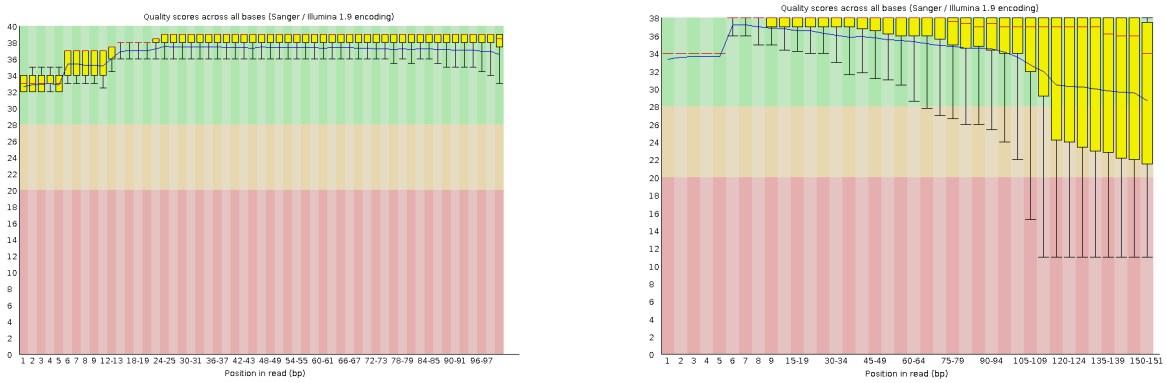
Illumina Sequencing Viewer

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.

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 (		



(a) hpx$_c sc_f astqc$ **(b)** tst15$_f astqc$

Figure 3: Comparison of Coverage Distributions per Amplicon

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

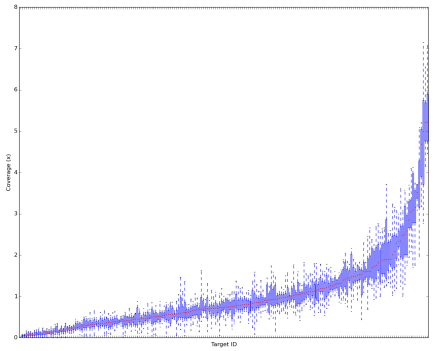
3.3 Coverage Analysis

Coverage Distribution TST15 vs Haloplex

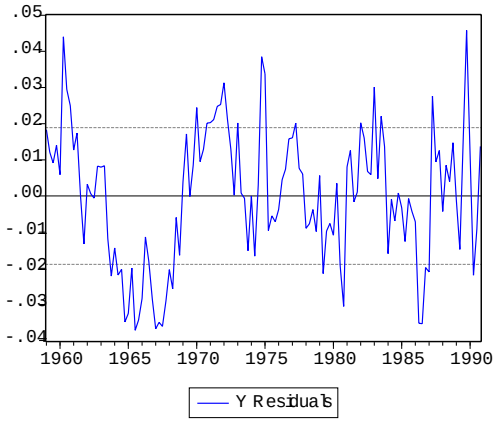
Coverage Distribution per Patient (check if correlation IQR with dCt)

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

Failed Amplicon Counter



(a) Agilent Haloplex ClearSeq Cancer



(b) Illumina TruSight Tumor 15

Figure 4: Comparison of Coverage Distributions per Amplicon

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?

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 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

[1] 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.

[2] T. S. Breusch and P. Schmidt, "Alternative forms of the Wald test: How long is a piece of string," *Communications in Statistics, Theory and Methods*, vol. 17, pp. 2789–2795, 1988.

[3] 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.