Doctoral Thesis

<Lung Precancer Analysis>

<Jaewoong Lee>

<Department of Biomedical Engineering>

Ulsan National Institute of Science and Technology

<2021>

<Lung Precancer Analysis>

<Jaewoong Lee>

<Department of Biomedical Engineering>

Ulsan National Institute of Science and Technology

Abstract

Contents

I Introduction 1

1.1 Lung Cancer 1

1.2 Non-small cell lung cancer 1

1.3 Lung Precancer 1

1.4 Study Objectives 1

II Materials 3

2.1 List of IPNs 3

2.2 Data Composition 3

III Methods 4

3.1 Workflows 4

IV Results 7

4.1 Quality Checks 7

4.2 Copy Number Variation Analyses 7

4.3 Somatic Short Variation Analyses 7

4.4 Variant Allele Frequency Analyses 7

4.5 Gene Fusion Analyses 7

4.6 Differences in Gene Expression levels 7

4.7 Bulk Cell Deconvolution Analyses 7

4.8 Mutational Signature Analyses 7

V Discussion 10

References 11

Acknowledgements 12

List of Figures

1 Common cancer survival rates (Hong et al., 2021) 2

2 Workflow for data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011) 5

3 Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011) 5

4 Germline short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011) 6

5 RNA-seq short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011) 6

6 FastQC results with WES data 8

7 FastQC results with WTS data 8

8 Comut Plot by LUSC 8

9 Comut Plot by LUAD 9

I Introduction

1.1 Lung Cancer

Lung cancer is the most common form of cancer as 12.3 % of all cancers (Minna, Roth, & Gazdar, 2002).

1.2 Non-small cell lung cancer

Lung adenocarcinoma (LUAD)

Lung squamous cell carcinoma (LUSC)

1.3 Lung Precancer

1.4 Study Objectives

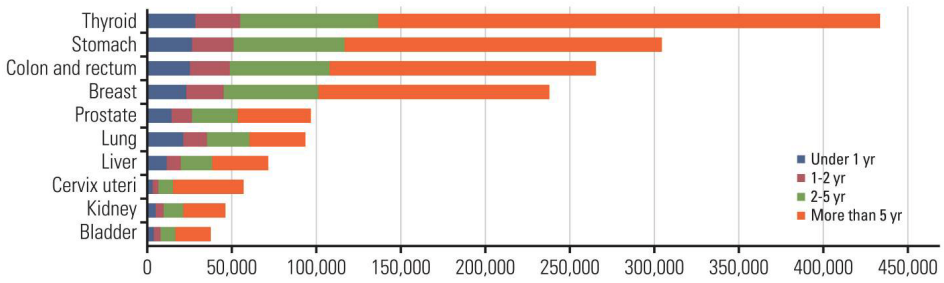


Figure 1: Common cancer survival rates (Hong et al., 2021)

II Materials

2.1 List of IPNs

Carcinoma *in situ*

Carcinoma *in situ* (CIS)

Adenocarcinoma *in situ*

Adenocarcinoma *in situ* (AIS)

Atypical Adenomatous Hyperplasia

Atypical adenomatous hyperplasia (AAH)

Dysplasia

Minimally Invasive Adenocarcinoma

Minimally invasive adenocarcinoma (MIA)

2.2 Data Composition

III Methods

3.1 Workflows

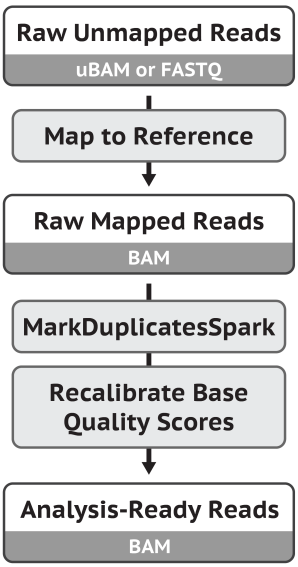


Figure 2: Workflow for data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

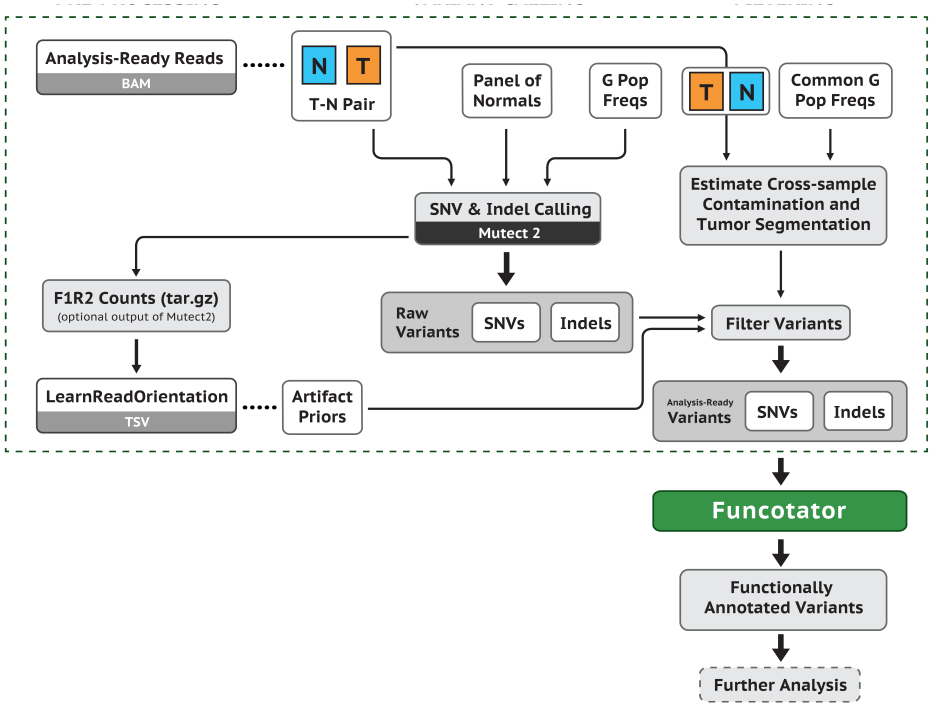


Figure 3: Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

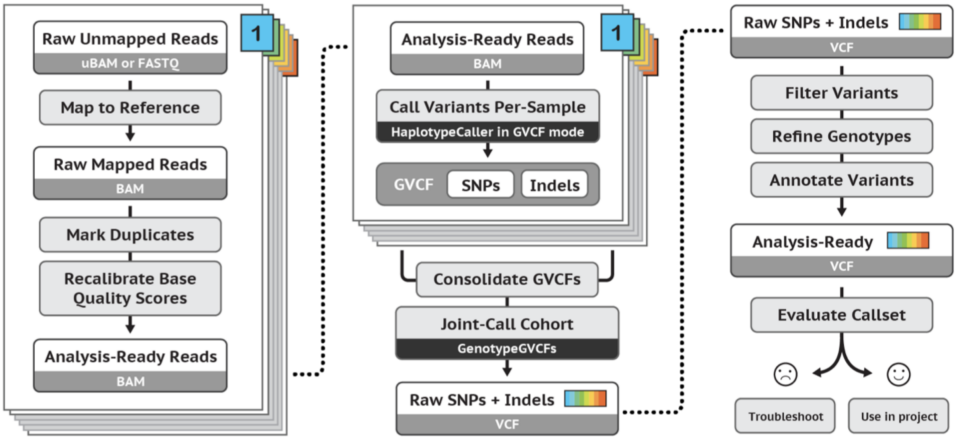


Figure 4: Germline short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

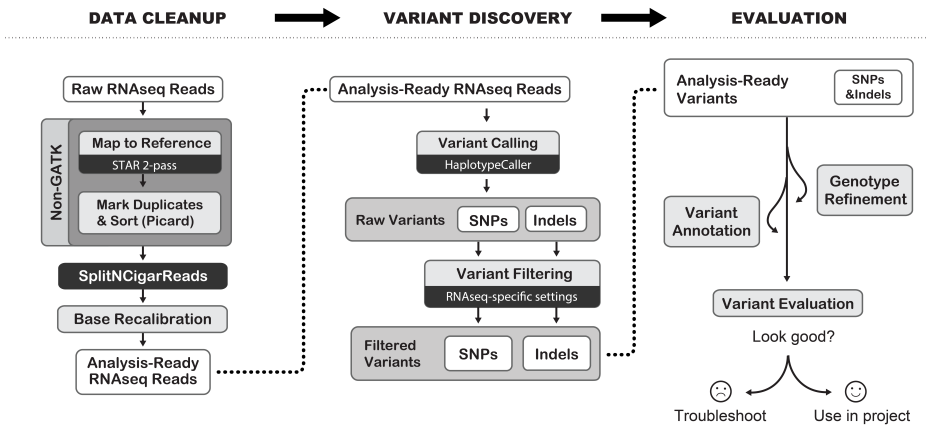


Figure 5: RNA-seq short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

IV Results

4.1 Quality Checks

Quality Checks with FastQC

Quality Checks with Picard

Findings in Quality Checks

4.2 Copy Number Variation Analyses

Purity and Ploidy

Copy Number Variation Plot

Findings in Copy Number Variation Analyses

4.3 Somatic Short Variation Analyses

Somatic Short Variation Analyses with Mutect2

Somatic Short Variant with Clinical Data

Findings in Somatic Short Variation Analyses

4.4 Variant Allele Frequency Analyses

Findings in Variant Allele Frequency Analyses

4.5 Gene Fusion Analyses

Findings in Gene Fusion Analyses

4.6 Differences in Gene Expression levels

4.7 Bulk Cell Deconvolution Analyses

Single-cell Reference Data

GSE131907 as Reference

GSE162498 as Reference

GSE179994 as Reference

Findings in Bulk Cell Deconvolution Analyses

4.8 Mutational Signature Analyses

Single Base Substitutions

Double Base Substitutions

Insertions and Deletions

Findings in Mutational Signature Analyses

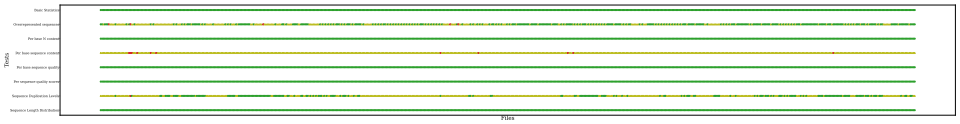


Figure 6: FastQC results with WES data

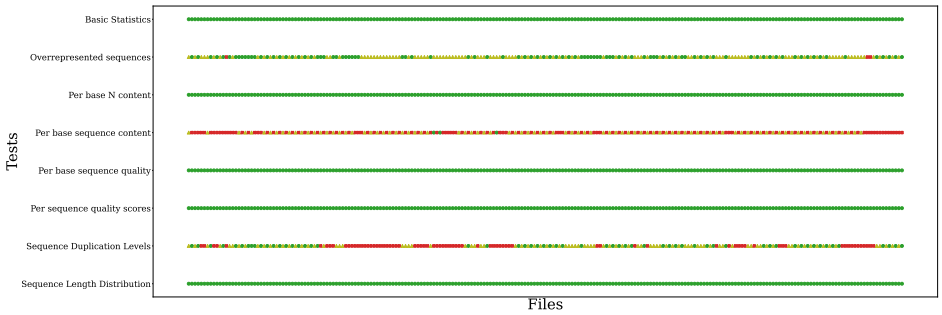
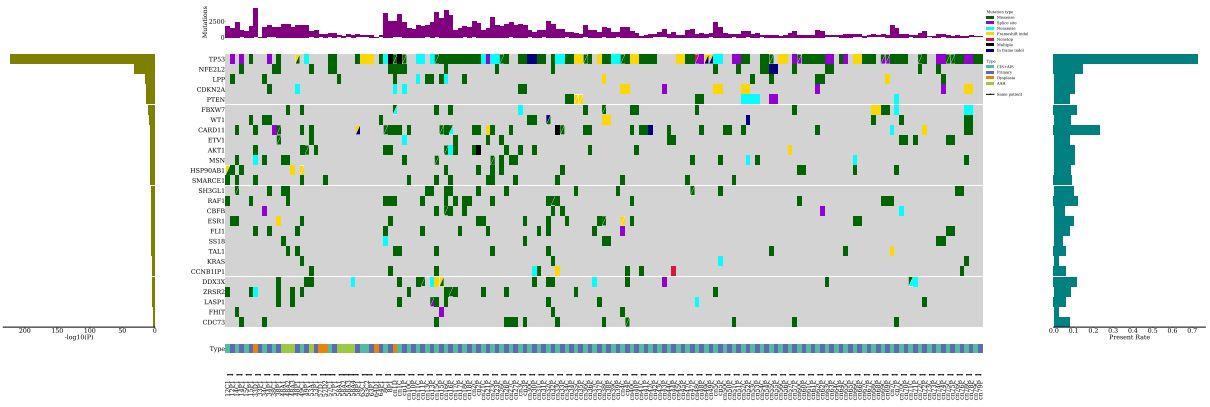
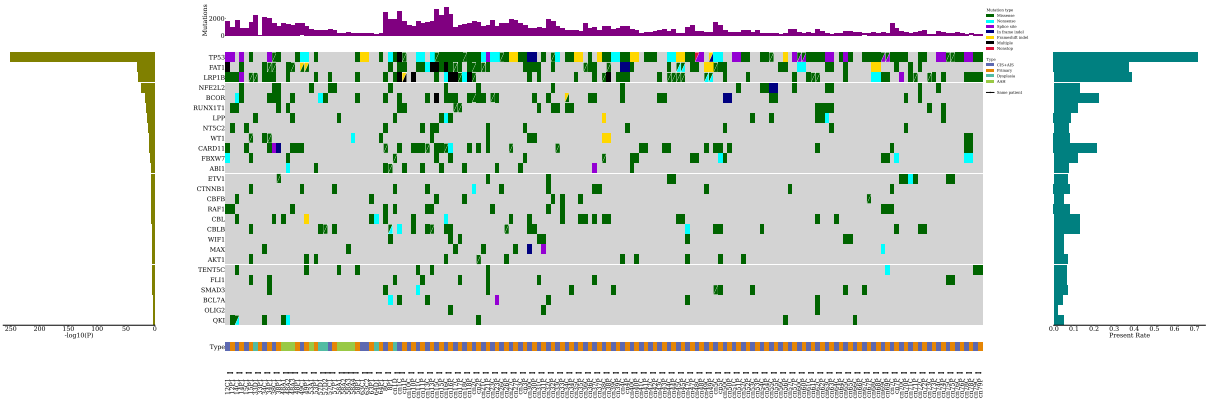


Figure 7: FastQC results with WTS data

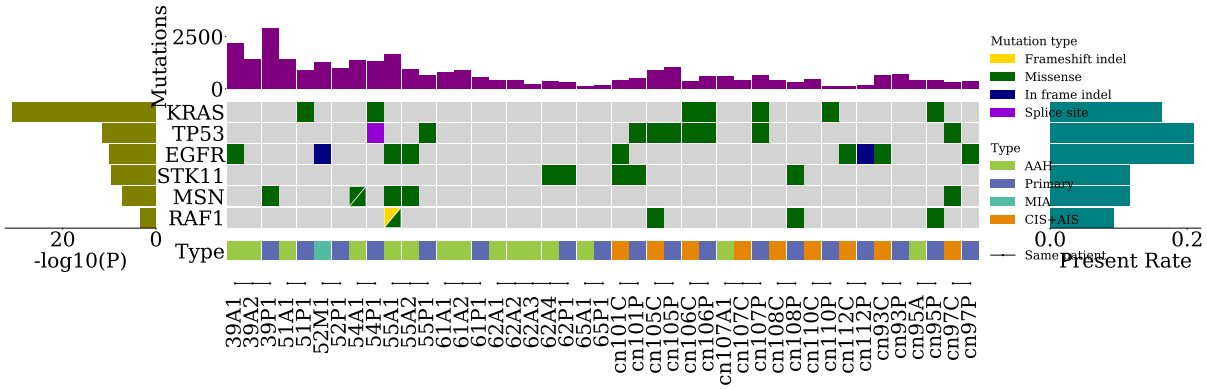


(a) BWA

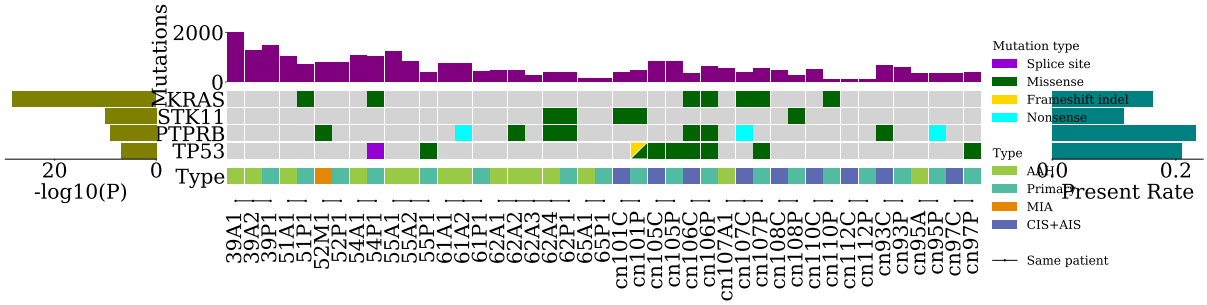


(b) Bowtie2

Figure 8: Comut Plot by LUSC



(a) BWA



(b) Bowtie2

Figure 9: Comut Plot by LUAD

V Discussion

References

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012, January). *FastQC*. Babraham Institute. Babraham, UK.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... others (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics*, *43*(5), 491.

Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., ... Eklund, A. C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, *26*(1), 64–70.

Hong, S., Won, Y.-J., Lee, J. J., Jung, K.-W., Kong, H.-J., Im, J.-S., ... others (2021). Cancer statistics in korea: Incidence, mortality, survival, and prevalence in 2018. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, *53*(2), 301.

Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on lung cancer. *Cancer cell*, *1*(1), 49–52.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... others (2013). From fastq data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11–10.

Acknowledgements

Thank you very much.