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

List of Figures

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

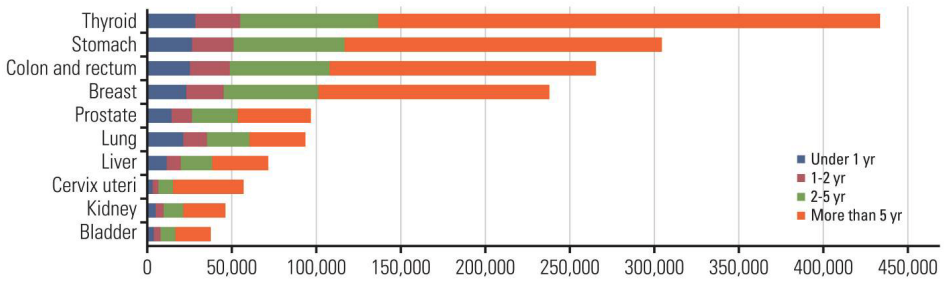


Figure 1: Common cancer survival rates (Hong et al., 2021)

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 Structure & Count

III Methods

3.1 Workflows

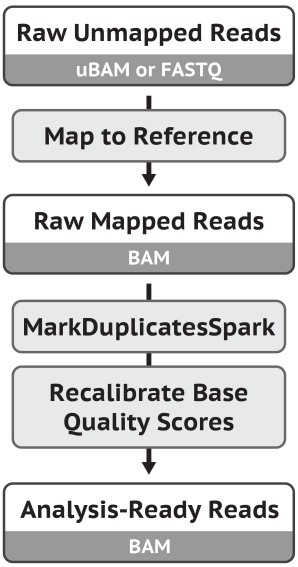


Figure 2: Workflow for data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

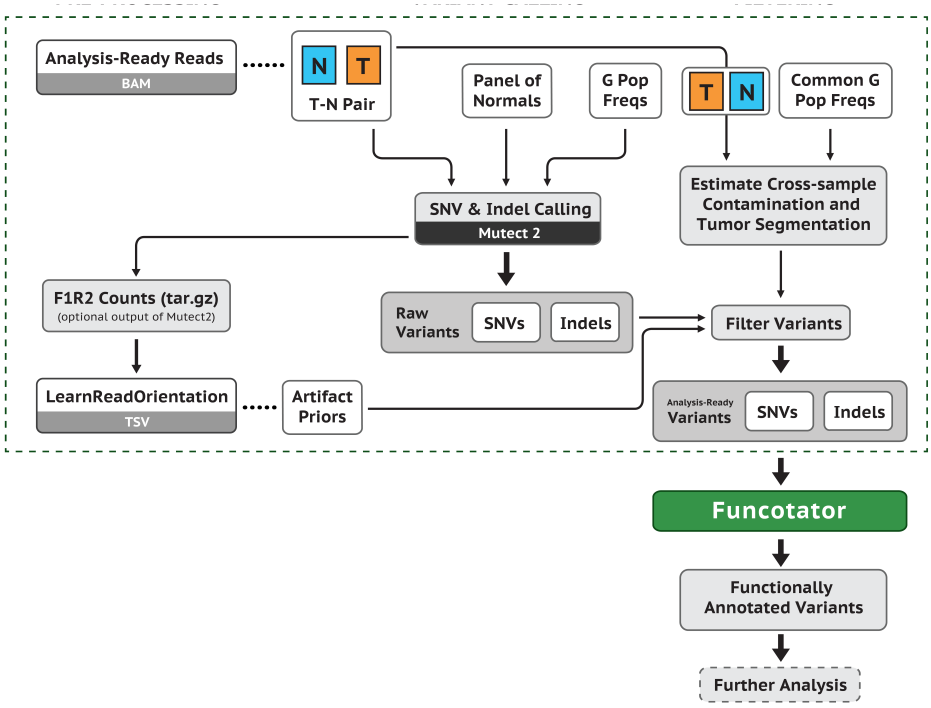


Figure 3: Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

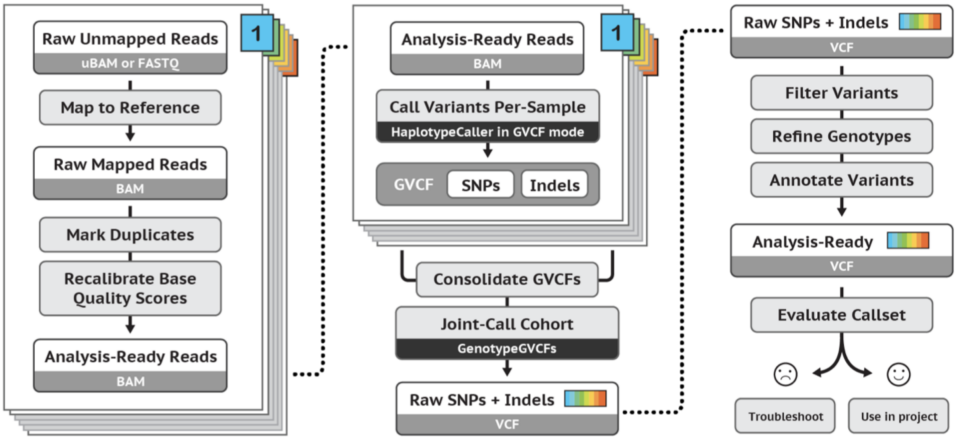


Figure 4: Germline short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

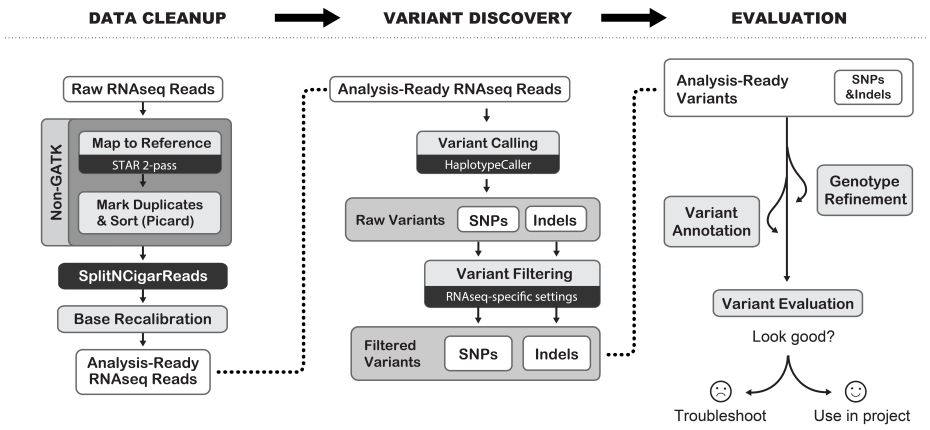


Figure 5: RNA-seq short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

IV Results

4.1 Quality Check

4.2 Quality Check with FastQC

Findings in Quality Check

4.3 Copy Number Variations

Copy Number Variation Analysis with Sequenza

Cellularities and Ploidies

Copy Number Variations

Findings in Copy Number Variation Analysis

4.4 Somatic Short Variation

Somatic Short Variation Analysis with Mutect2

Findings in Somatic Short Variation Analysis

4.5 Variant Allele Frequencies

4.6 Differences in Gene Expression levels

4.7 Bulk Cell Deconvolution

Single-cell Reference Data

CIBERSORTx

BisqueRNA

MuSiC

SCDC

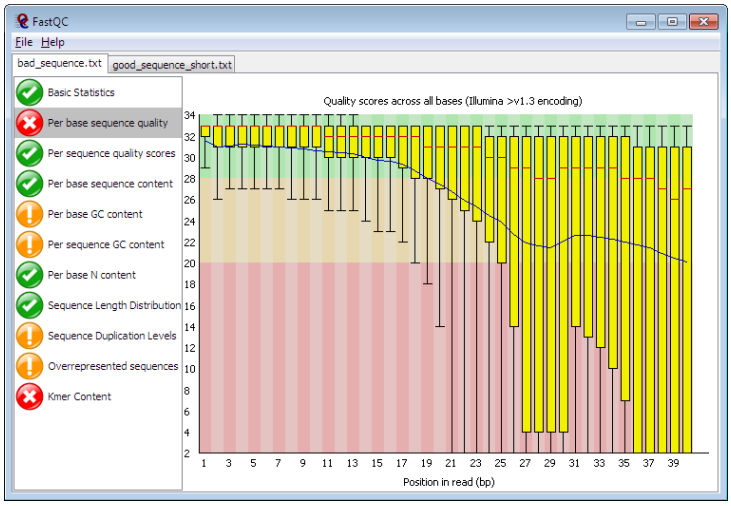


Figure 6: Example of FastQC Result (Andrews et al., 2012)

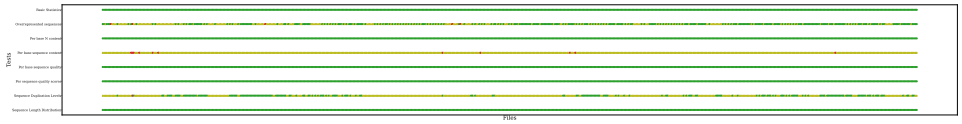


Figure 7: FastQC results with WES data

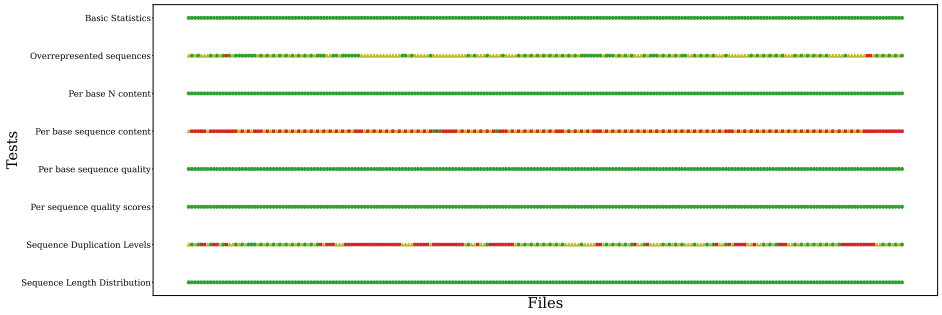


Figure 8: FastQC results with WTS data

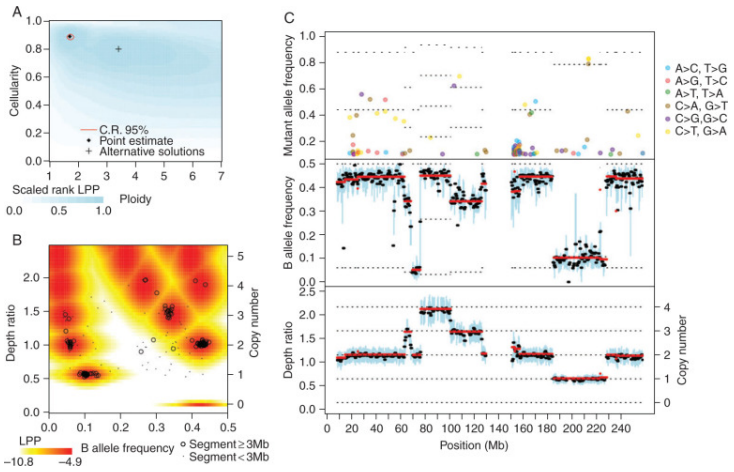


Figure 9: Representative Output of the Sequenza (Favero et al., 2015)

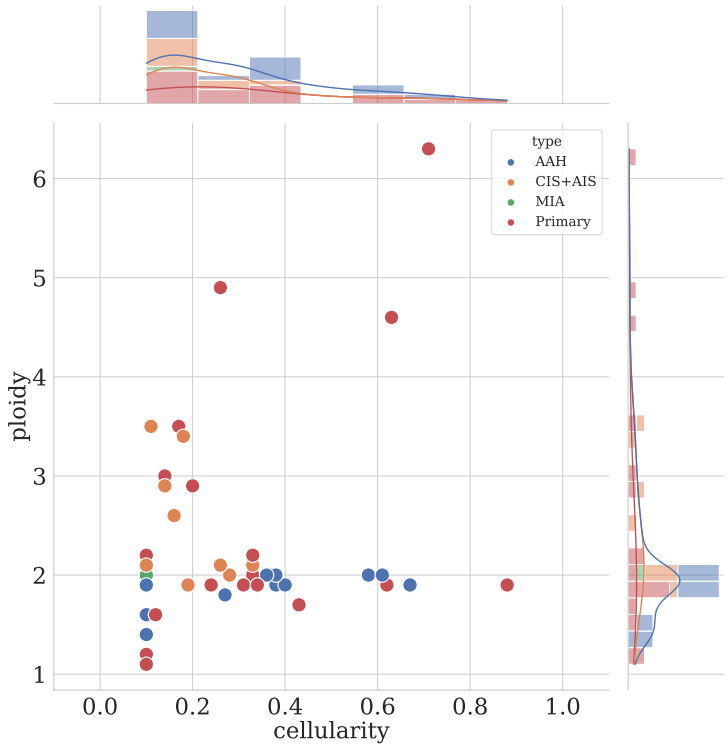


Figure 10: Cellularities and Ploidies by BWA in ADC

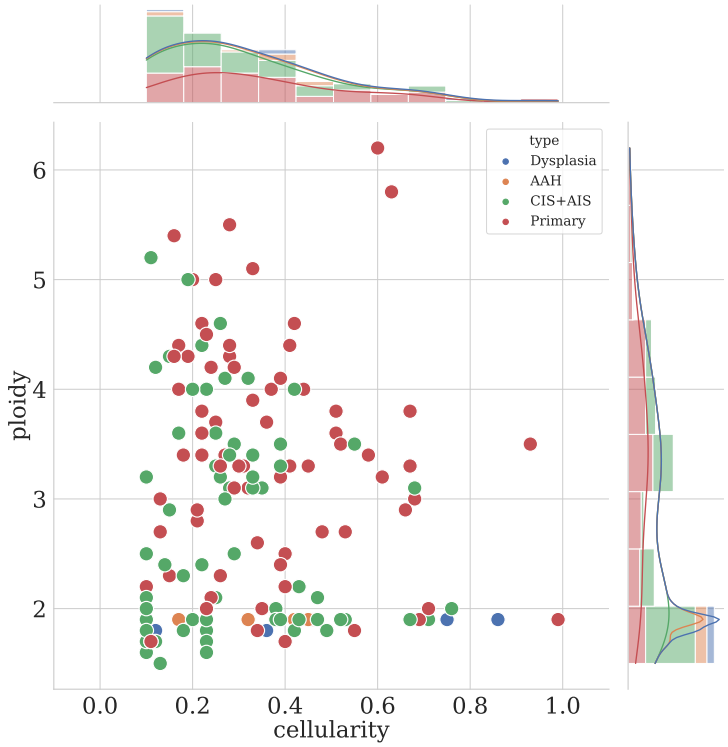


Figure 11: Cellularities and Ploidies by BWA in SQC

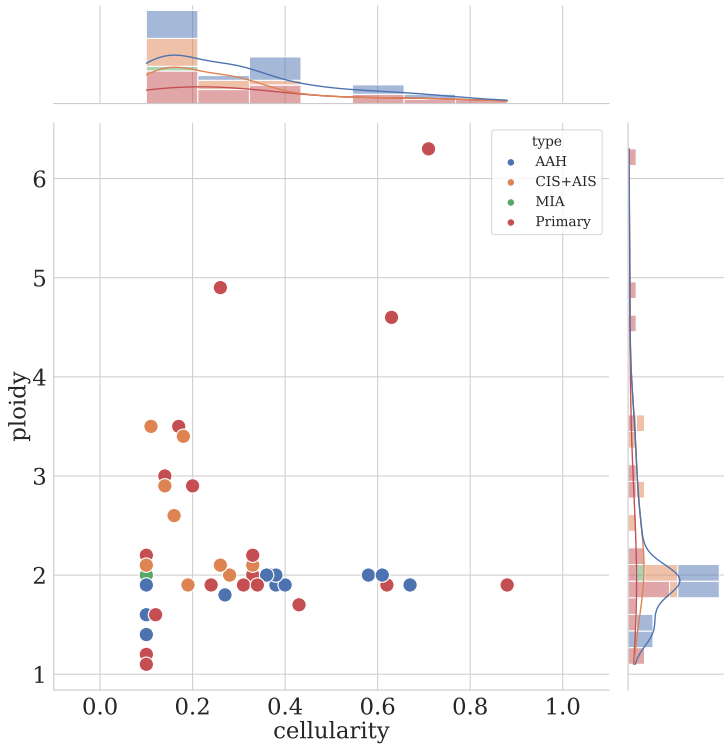


Figure 12: Cellularities and Ploidies by Bowtie2 in ADC

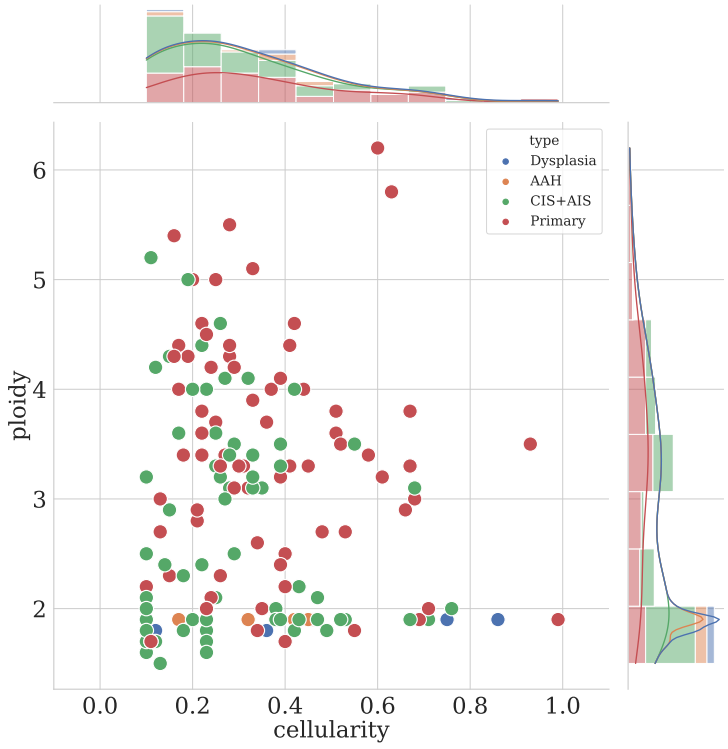


Figure 13: Cellularities and Ploidies by Bowtie2 in SQC

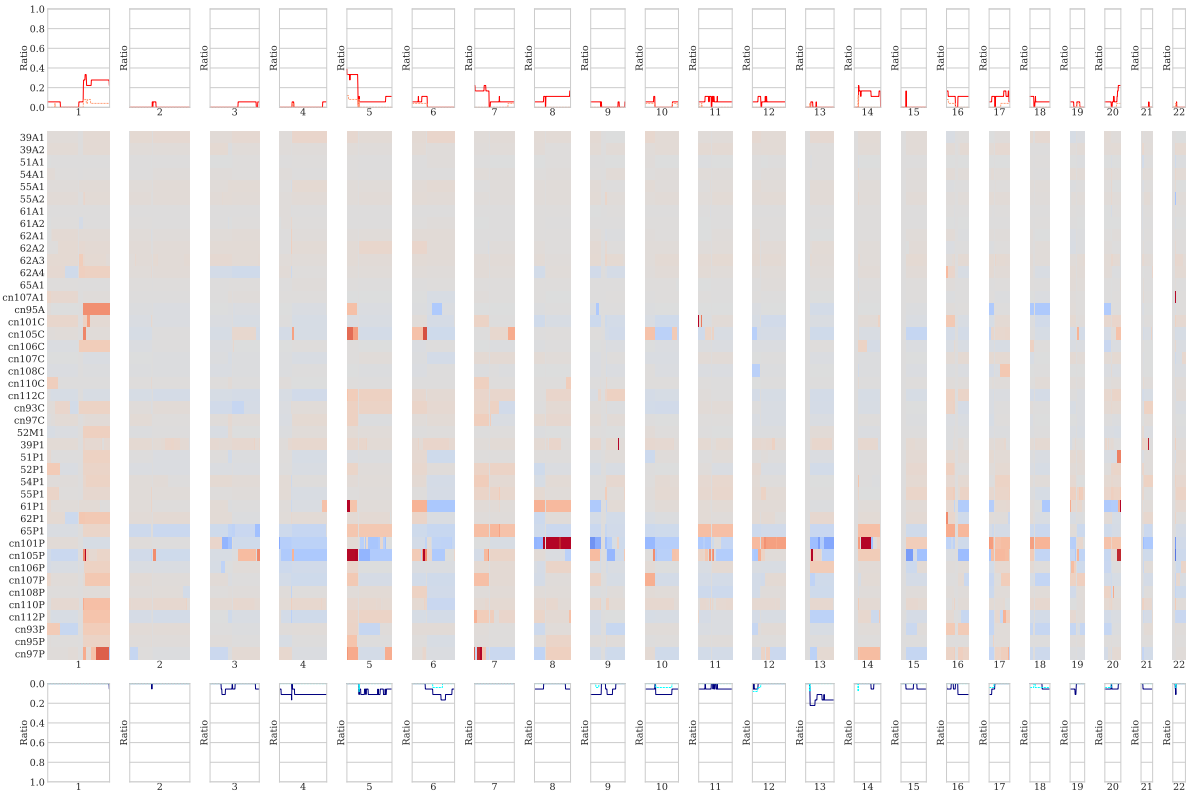


Figure 14: CNV plot by BWA in ADC

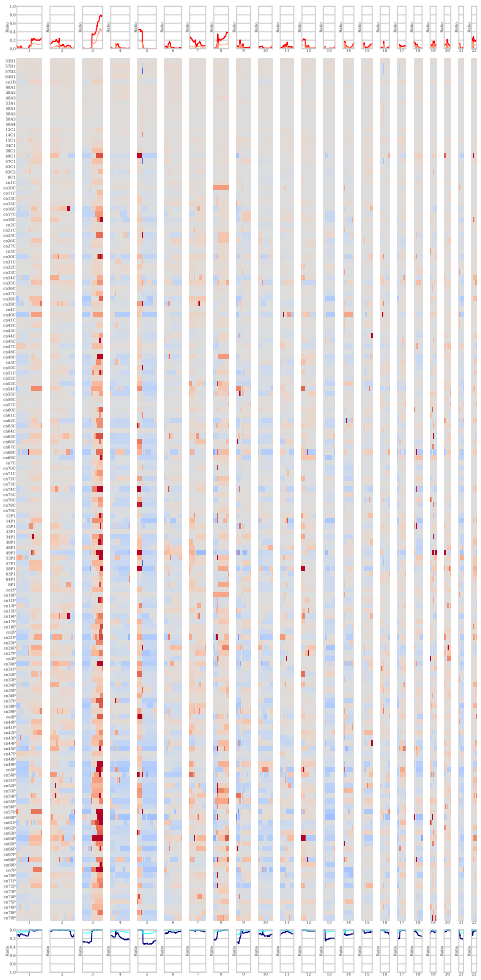


Figure 15: CNV plot by BWA in SQC

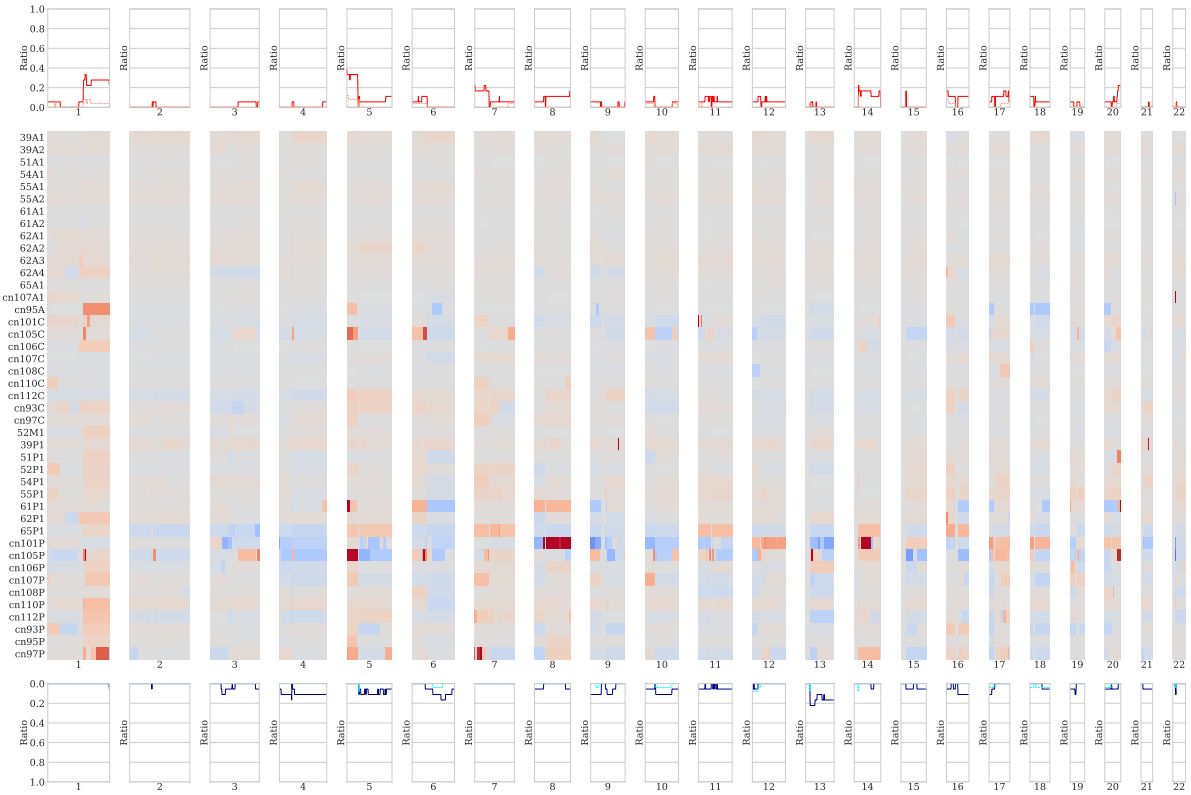


Figure 16: CNV plot by Bowtie2 in ADC

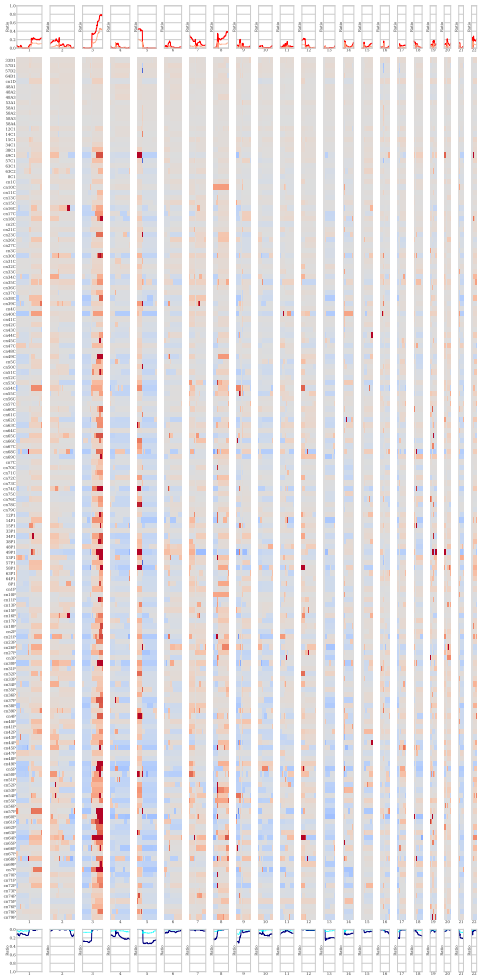


Figure 17: CNV plot by Bowtie2 in SQC

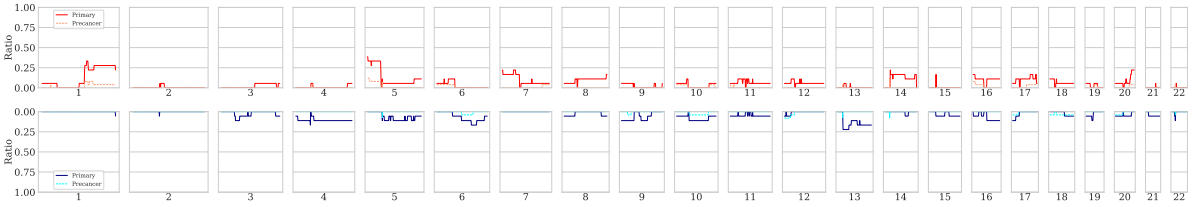


Figure 18: Simple CNV plot by BWA in ADC

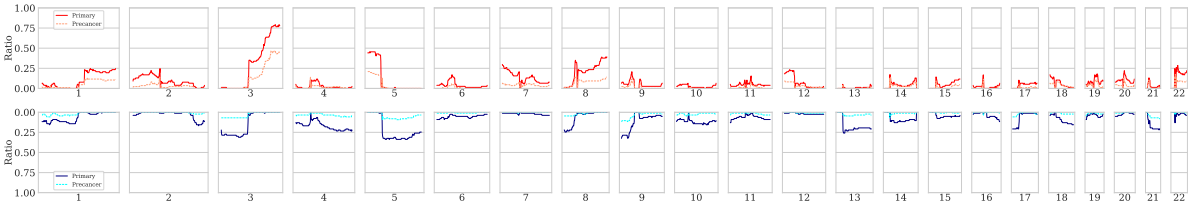


Figure 19: Simple CNV plot by BWA in SQC

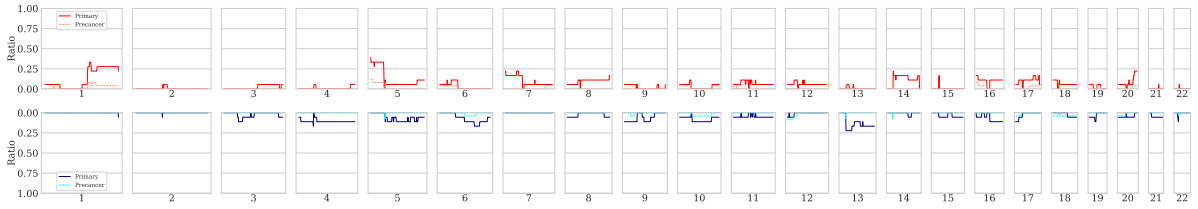


Figure 20: Simple CNV plot by Bowtie2 in ADC

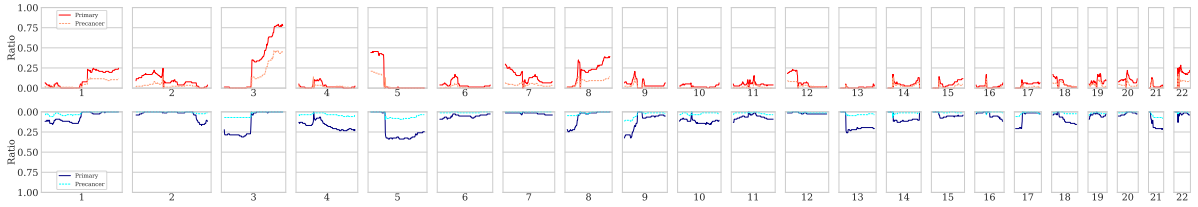


Figure 21: Simple CNV plot by Bowtie2 in SQC

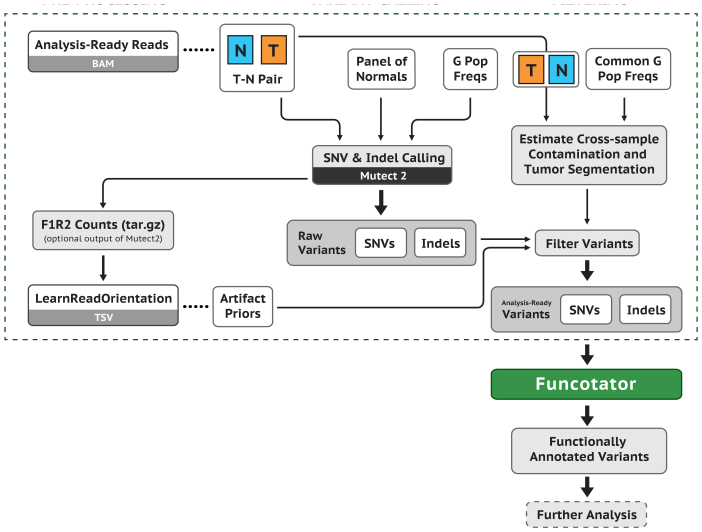


Figure 22: Somatic Short Variant Discovery Workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

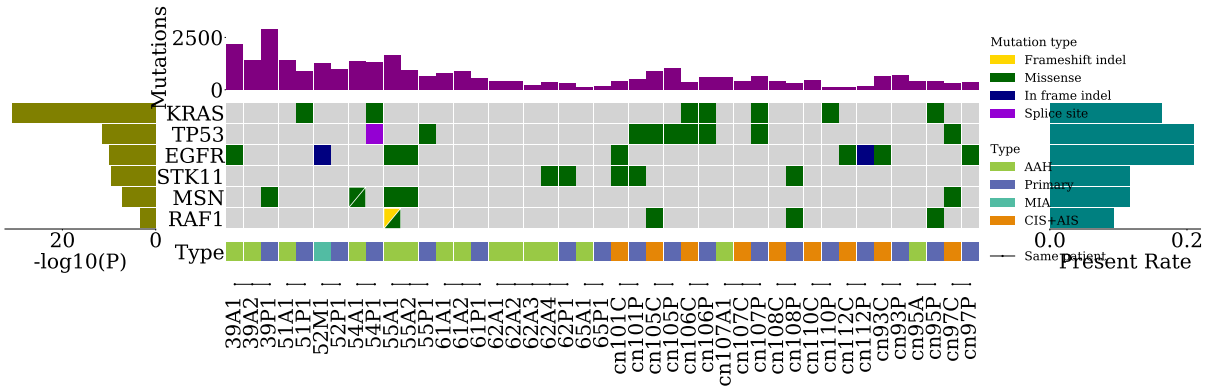


Figure 23: CoMut plot by BWA in ADC

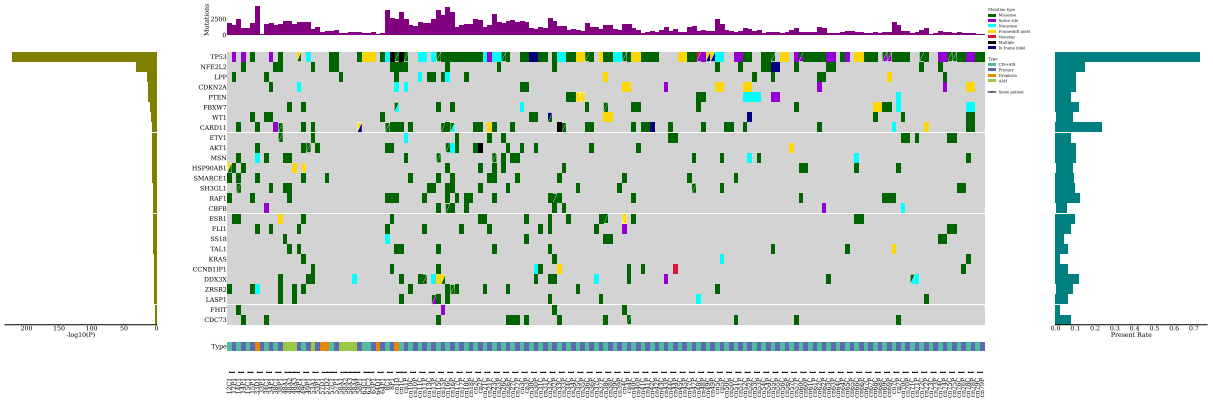


Figure 24: CoMut plot by BWA in SQC

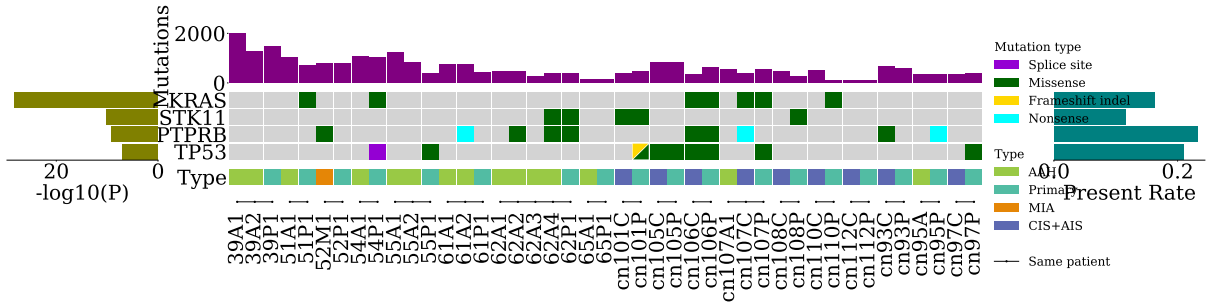


Figure 25: CoMut plot by Bowtie2 in ADC

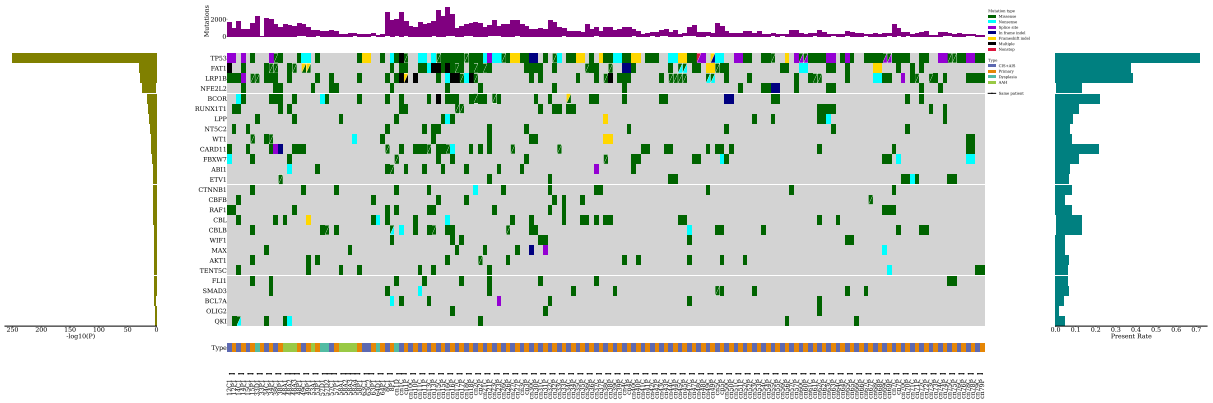
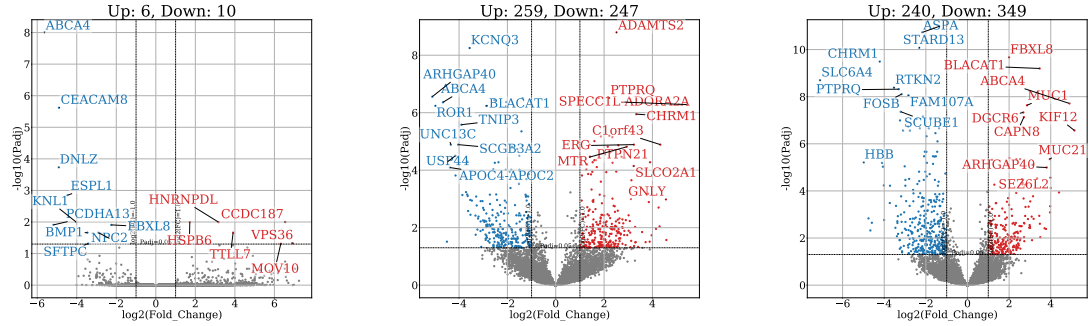


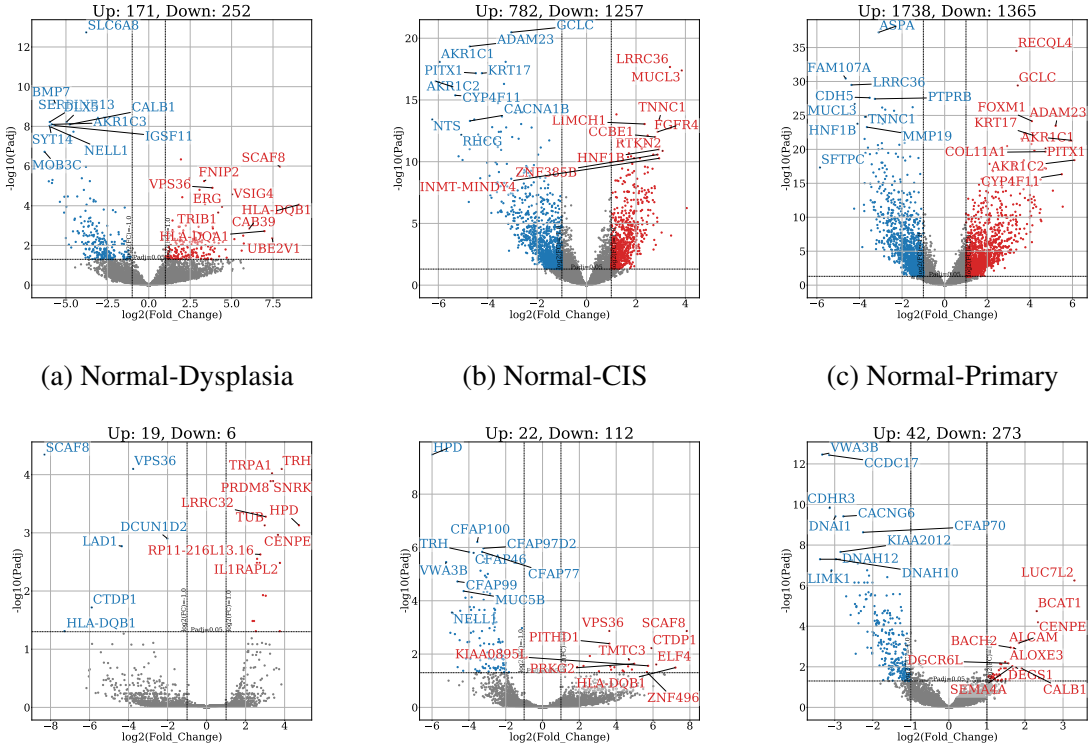
Figure 26: CoMut plot by Bowtie2‘ in SQC



(a) Normal-AAH (b) Normal-AIS (c) Normal-Primary

(d) AAH-AIS (e) AAH-Primary (f) AIS-Primary

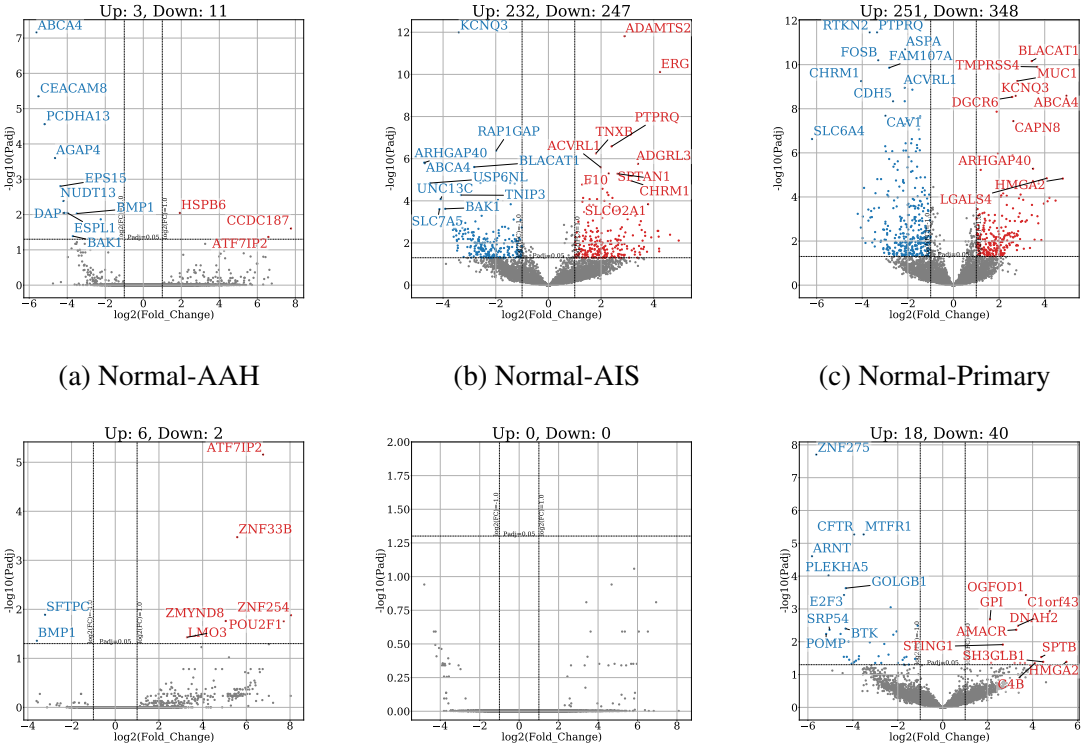
Figure 27: DEG volcano plots by Bowtie2 in ADC



(a) Normal-Dysplasia (b) Normal-CIS (c) Normal-Primary

(d) Dysplasia-CIS (e) Dysplasia-Primary (f) CIS-Primary

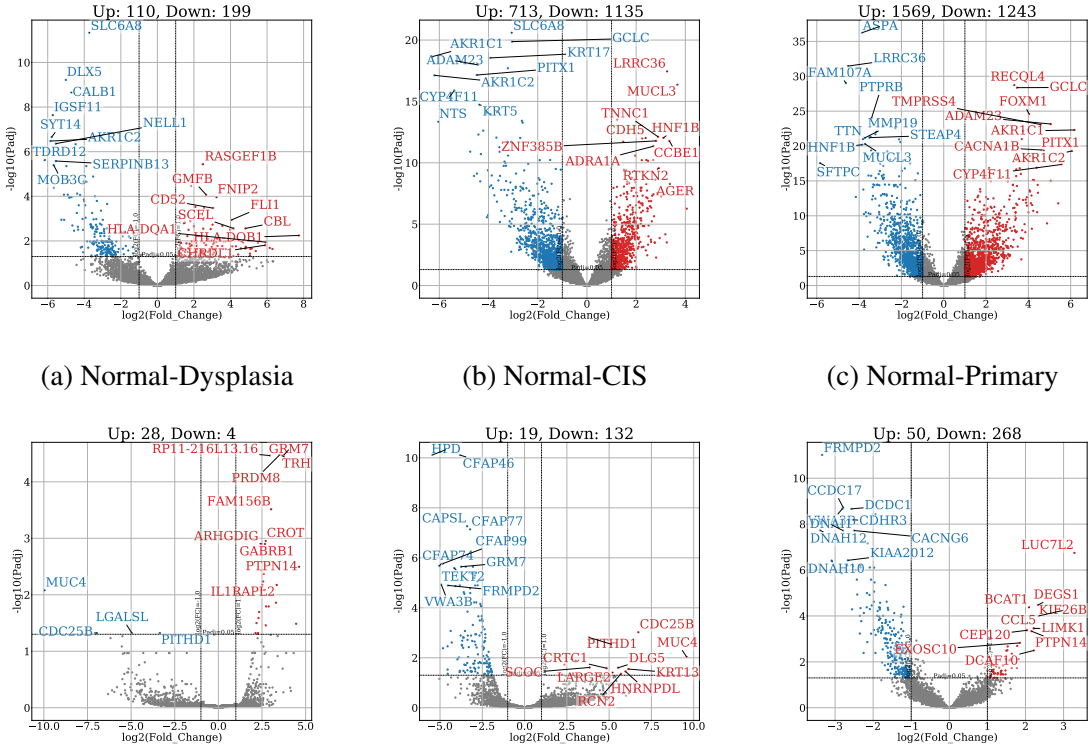
Figure 28: DEG volcano plots by Bowtie2 in SQC



(a) Normal-AAH (b) Normal-AIS (c) Normal-Primary

(d) AAH-AIS (e) AAH-Primary (f) AIS-Primary

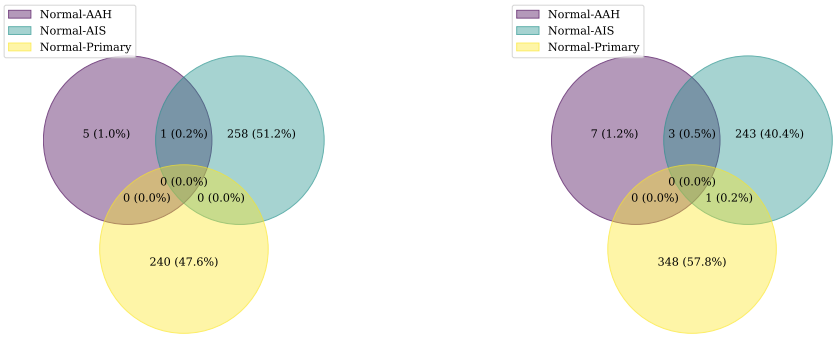
Figure 29: DEG volcano plots by STAR in ADC



(a) Normal-Dysplasia (b) Normal-CIS (c) Normal-Primary

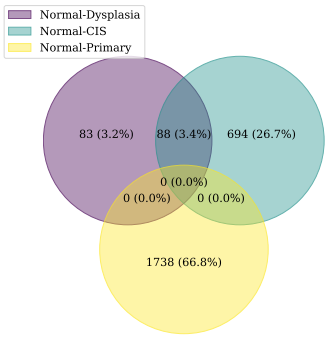
(d) Dysplasia-CIS (e) Dysplasia-Primary (f) CIS-Primary

Figure 30: DEG volcano plots by Bowtie2 in SQC

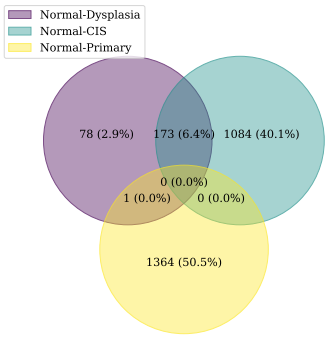


(a) Up-regulated (b) Down-regulated

Figure 31: DEG Venn Diagram by Bowtie2 in ADC

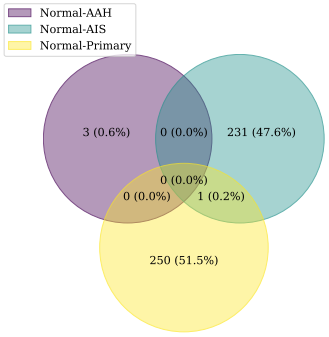


(a) Up-regulated

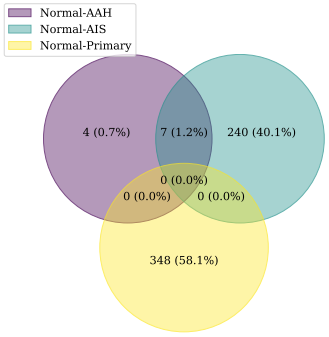


(b) Down-regulated

Figure 32: DEG Venn Diagram by Bowtie2 in SQC

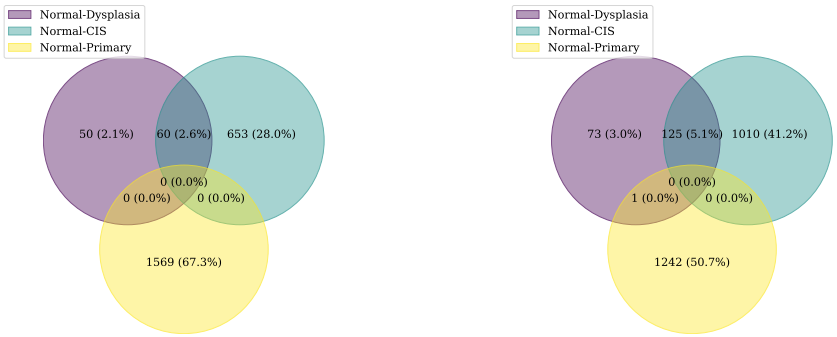


(a) Up-regulated



(b) Down-regulated

Figure 33: DEG Venn Diagram by STAR in ADC



(a) Up-regulated (b) Down-regulated

Figure 34: DEG Venn Diagram by STAR in SQC

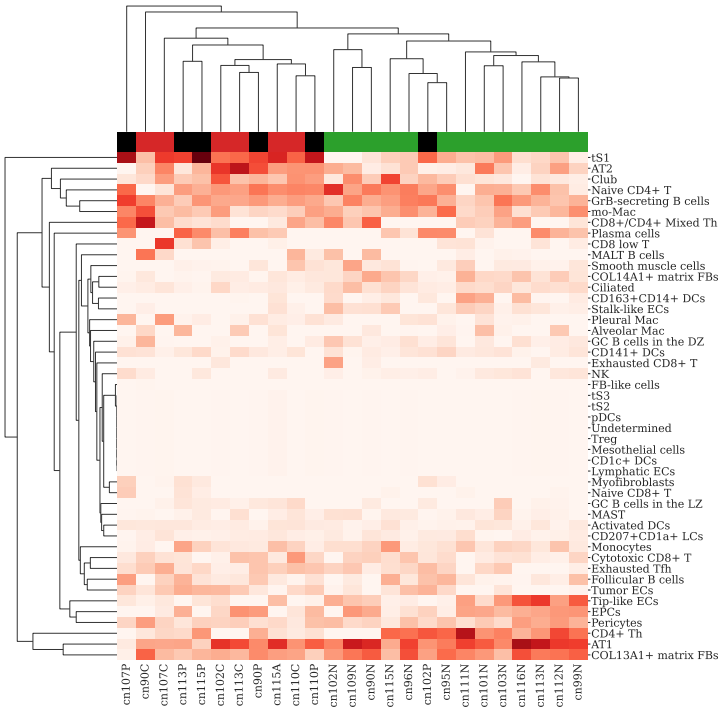


Figure 35: Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in ADC

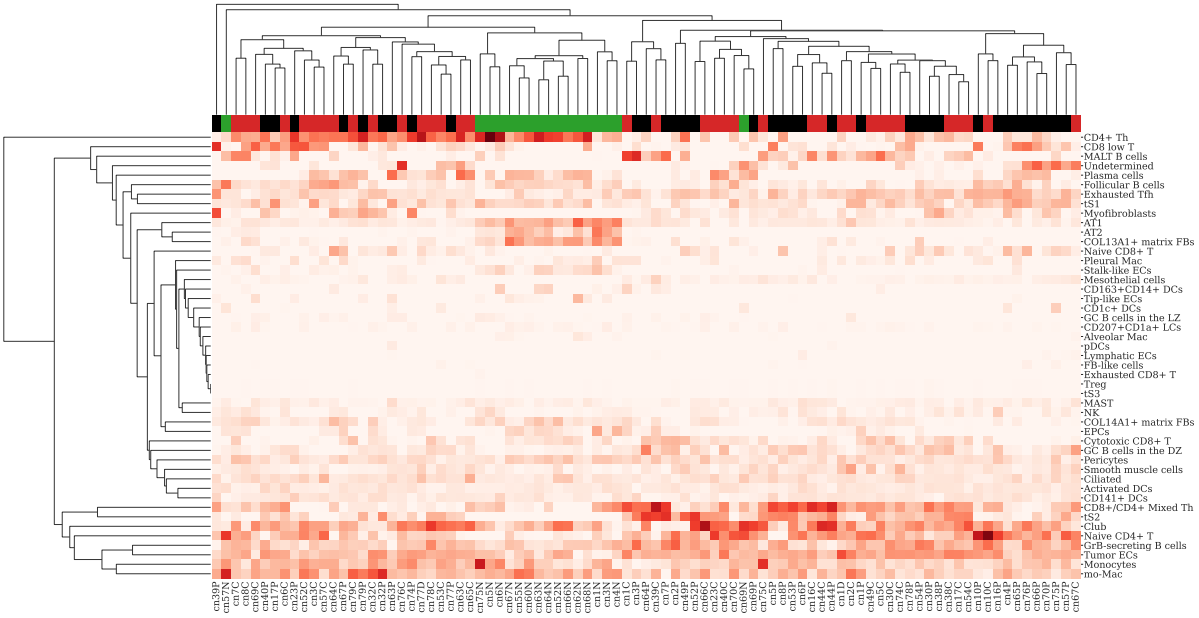


Figure 36: Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in SQC

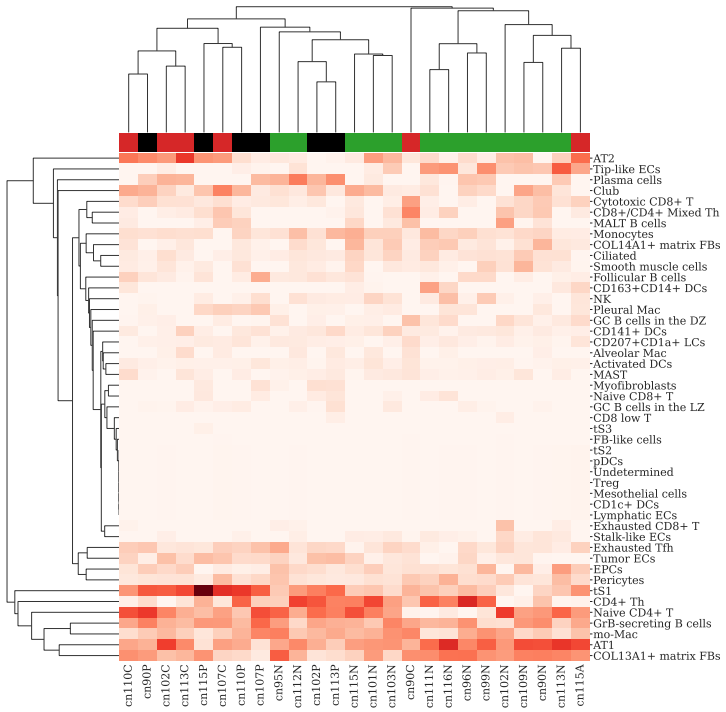


Figure 37: Cell deconvolution clustermap by STAR and CIBERSORTx in ADC

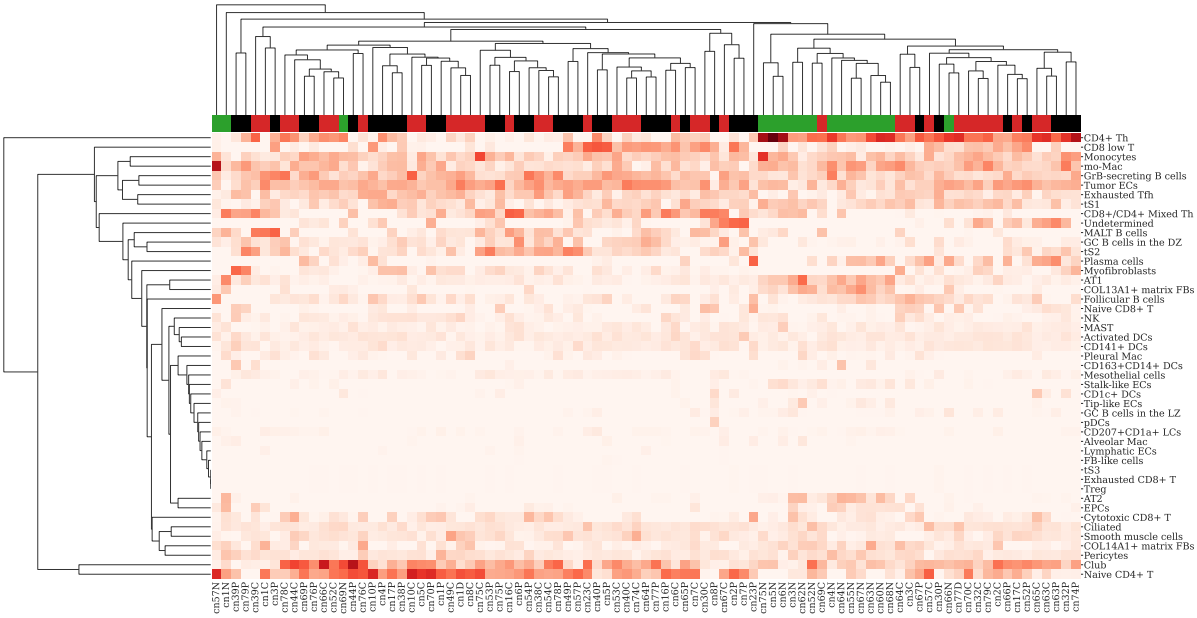


Figure 38: Cell deconvolution clustermap by STAR and CIBERSORTx in SQC

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.