Doctoral Thesis

<Lung Precancer Analysis>

<Jaewoong Lee>

<Department of Biomedical Engineering>

Ulsan National Institute of Science and Technology

<2023>

<Lung Precancer Analysis>

<Jaewoong Lee>

<Department of Biomedical Engineering>

Ulsan National Institute of Science and Technology

Abstract

Contents

List of Figures

List of Tables

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)

LUSC vs. LUAD

1.3 Lung Precancer

1.4 Study Objectives

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

Table 1: WES Data Composition

Cancer Subtype	Stage	Number of Samples
LUSC	Normal	77
	Dysplasia	5
	AAH	8
	CIS+AIS	73
	Primary	77
	Total	240
LUAD	Normal	18
	AAH	15
	CIS+AIS	9
	MIA	1
	Primary	18
	Total	61

Table 2: WTS Data Composition

Cancer Subtype	Stage	Number of Samples
LUSC	Normal	17
	Dysplasia	2
	CIS+AIS	34
	Primary	36
	Total	89
LUAD	Normal	13
	AAH	1
	CIS+AIS	5
	Primary	6
	Total	25

III Methods

IV Results

4.1 Quality Checks

Quality Checks with FastQC

Quality Checks with Depths

Quality Checks with Picard

Findings in Quality Checks

4.2 Copy Number Variation Analyses

Purity and Ploidy

Copy Number Variation Analyses

Copy Number Variation Analyses with Recurrence

Copy Number Variation Analyses with Smoking History

Gistic Analyses

Gistic Analyses with Recurrence

Gistic Analyses with Smoking History

Findings in Copy Number Variation Analyses

4.3 Single Nucleotide 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 Bulk Cell Deconvolution Analyses

Single-cell Reference Data

GSE131907 as Reference

GSE162498 as Reference

GSE179994 as Reference

Findings in Bulk Cell Deconvolution Analyses

4.6 Mutational Signature Analyses

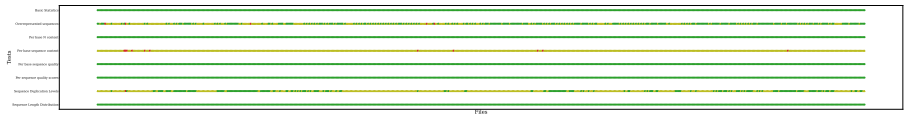


Figure 1: FastQC results with WES data

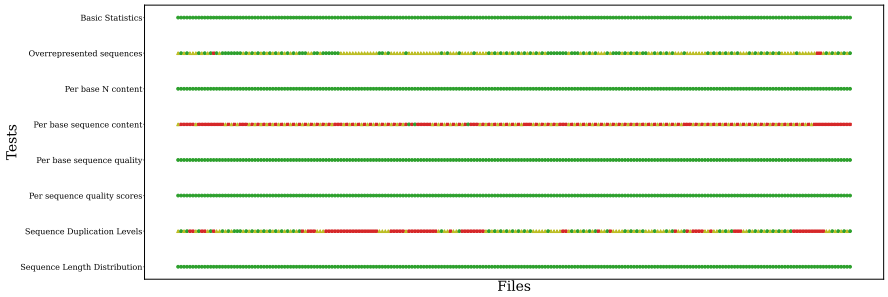
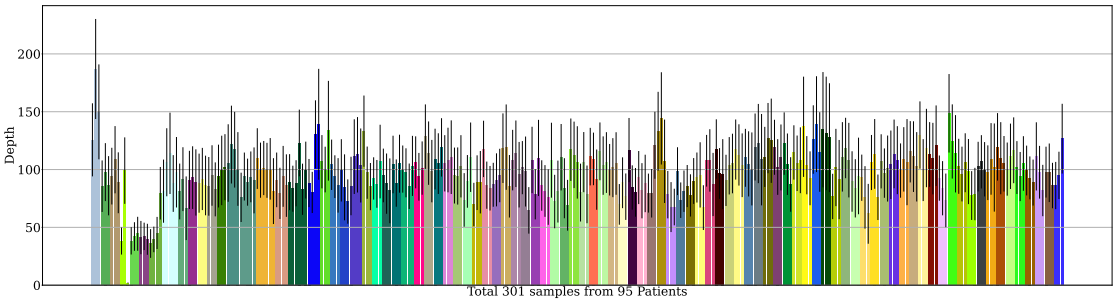
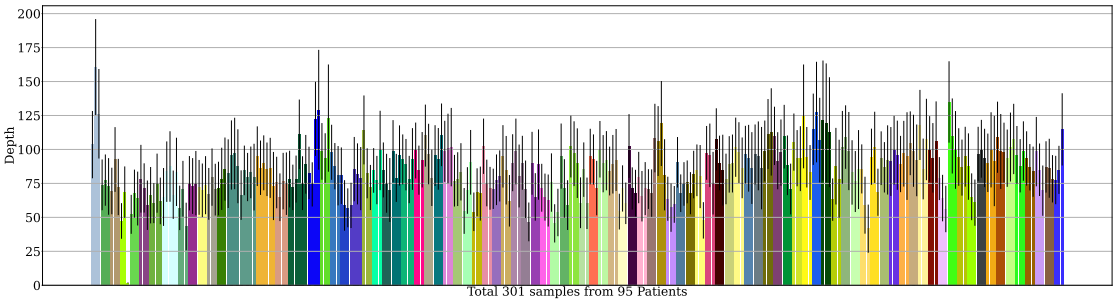


Figure 2: FastQC results with WTS data

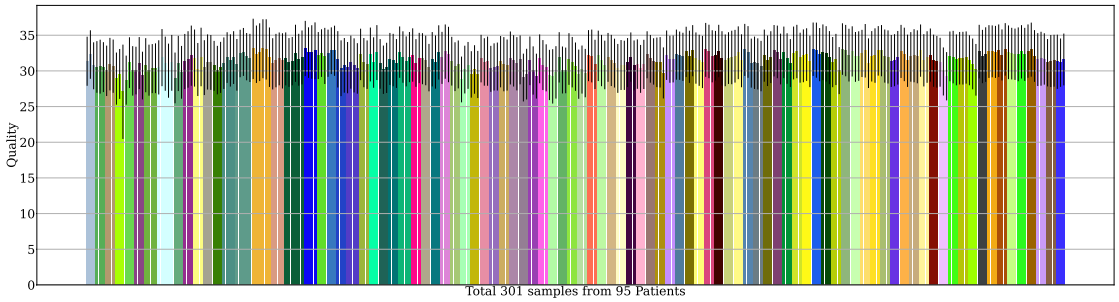


(a) BWA

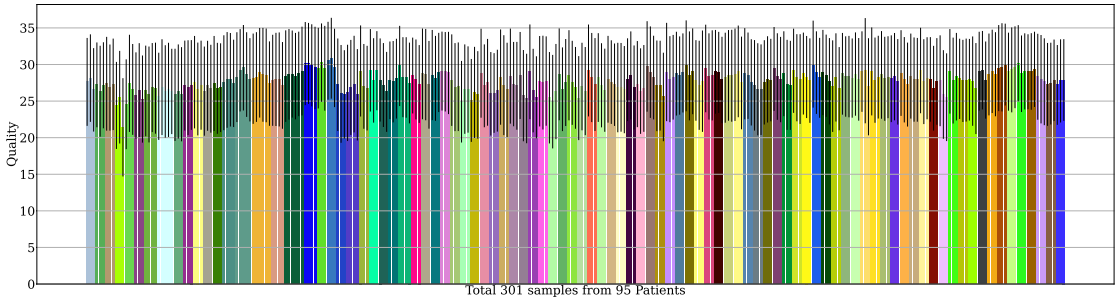


(b) Bowtie2

Figure 3: Depths plot with WES data

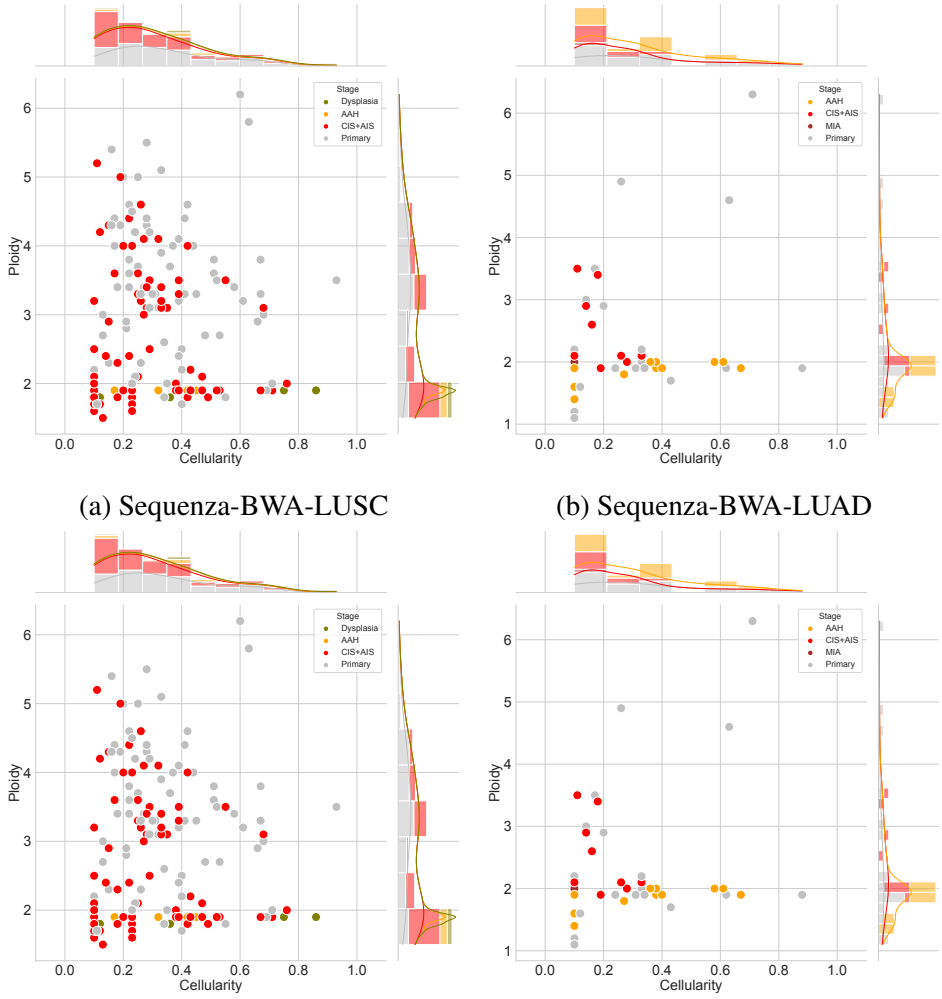


(a) BWA



(b) Bowtie2

Figure 4: Quality Distribution by Samples



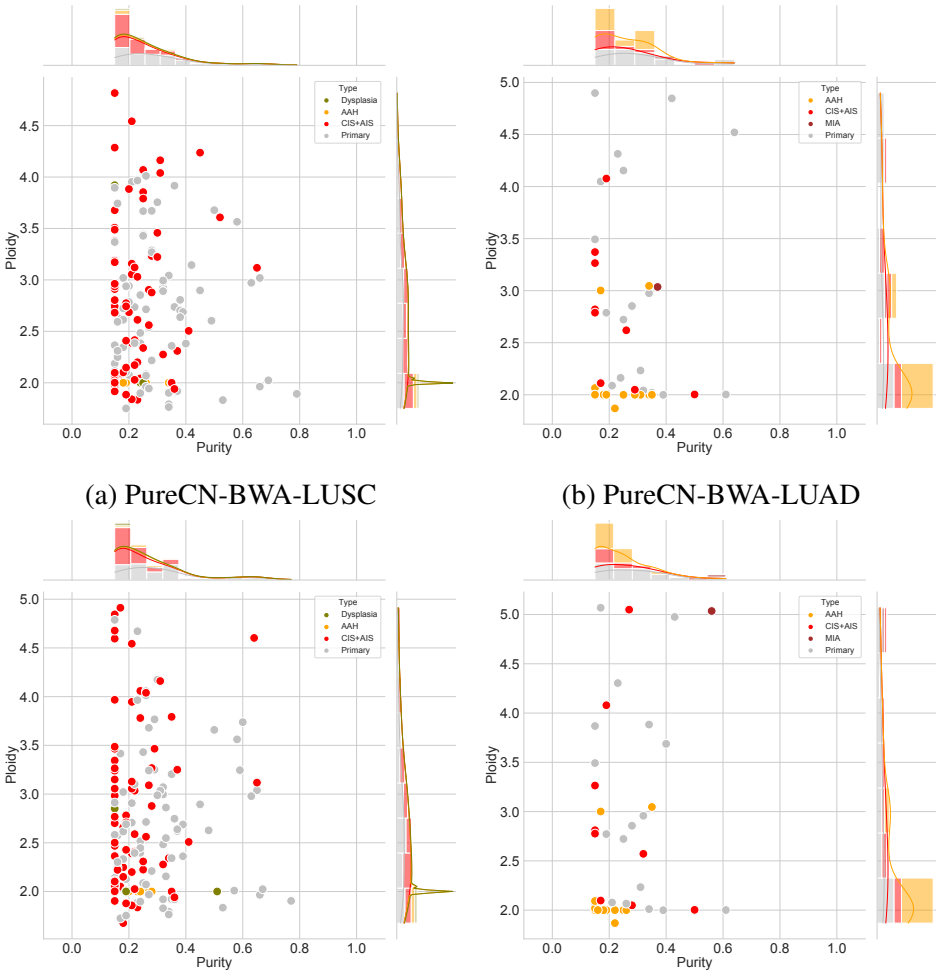
(a) Sequenza-BWA-LUSC

(b) Sequenza-BWA-LUAD

(c) Sequenza-Bowtie2-LUSC

(d) Sequenza-Bowtie2-LUAD

Figure 5: Sequenza Cellularity and Ploidy Plots



(a) PureCN-BWA-LUSC

(b) PureCN-BWA-LUAD

(c) PureCN-Bowtie2-LUSC

(d) PureCN-Bowtie2-LUAD

Figure 6: PureCN Purity and Ploidy Plots

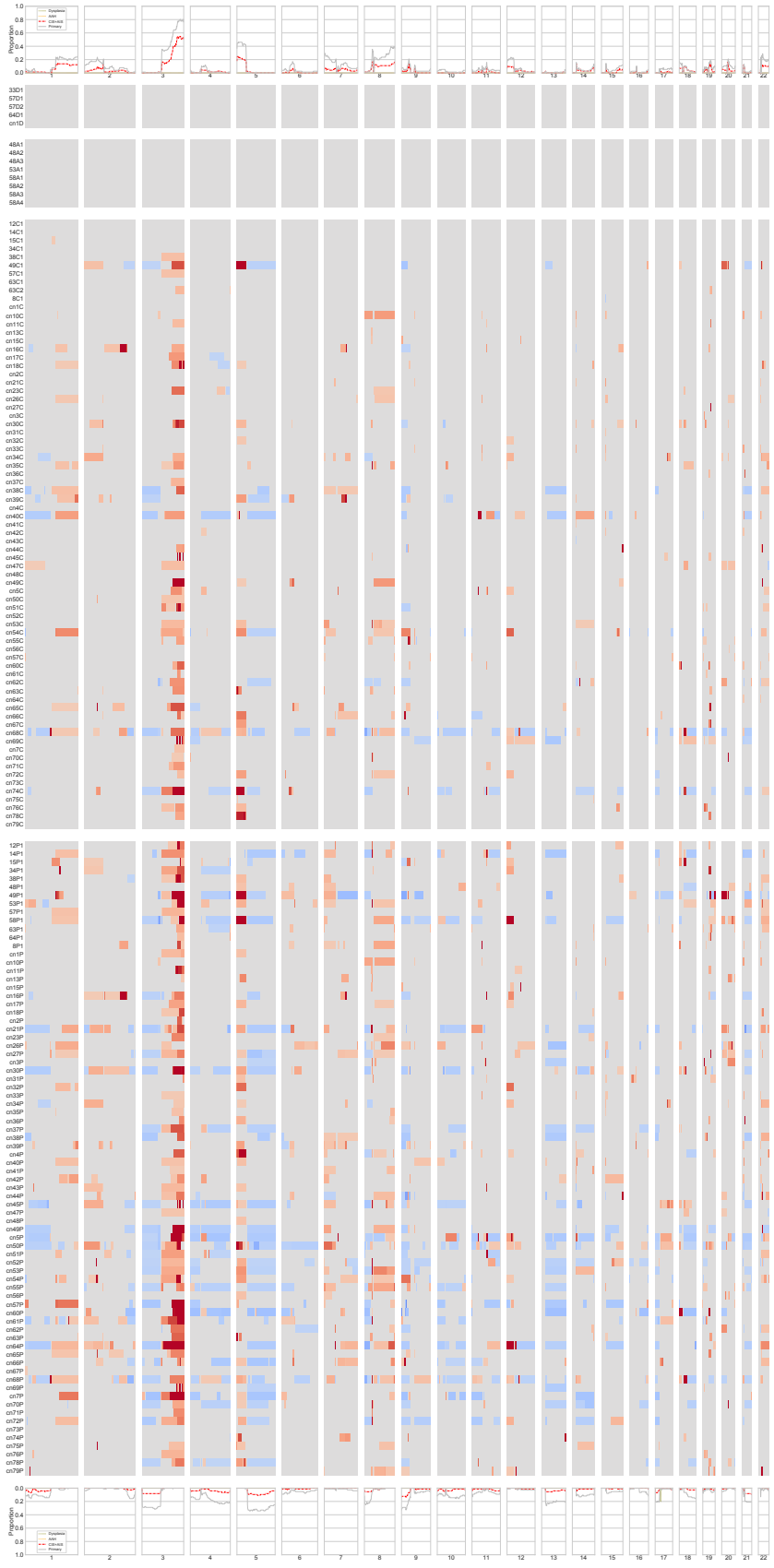


Figure 7: Sequenza LUSC Genome View Plot

Figure 8: PureCN LUSC Genome View Plot

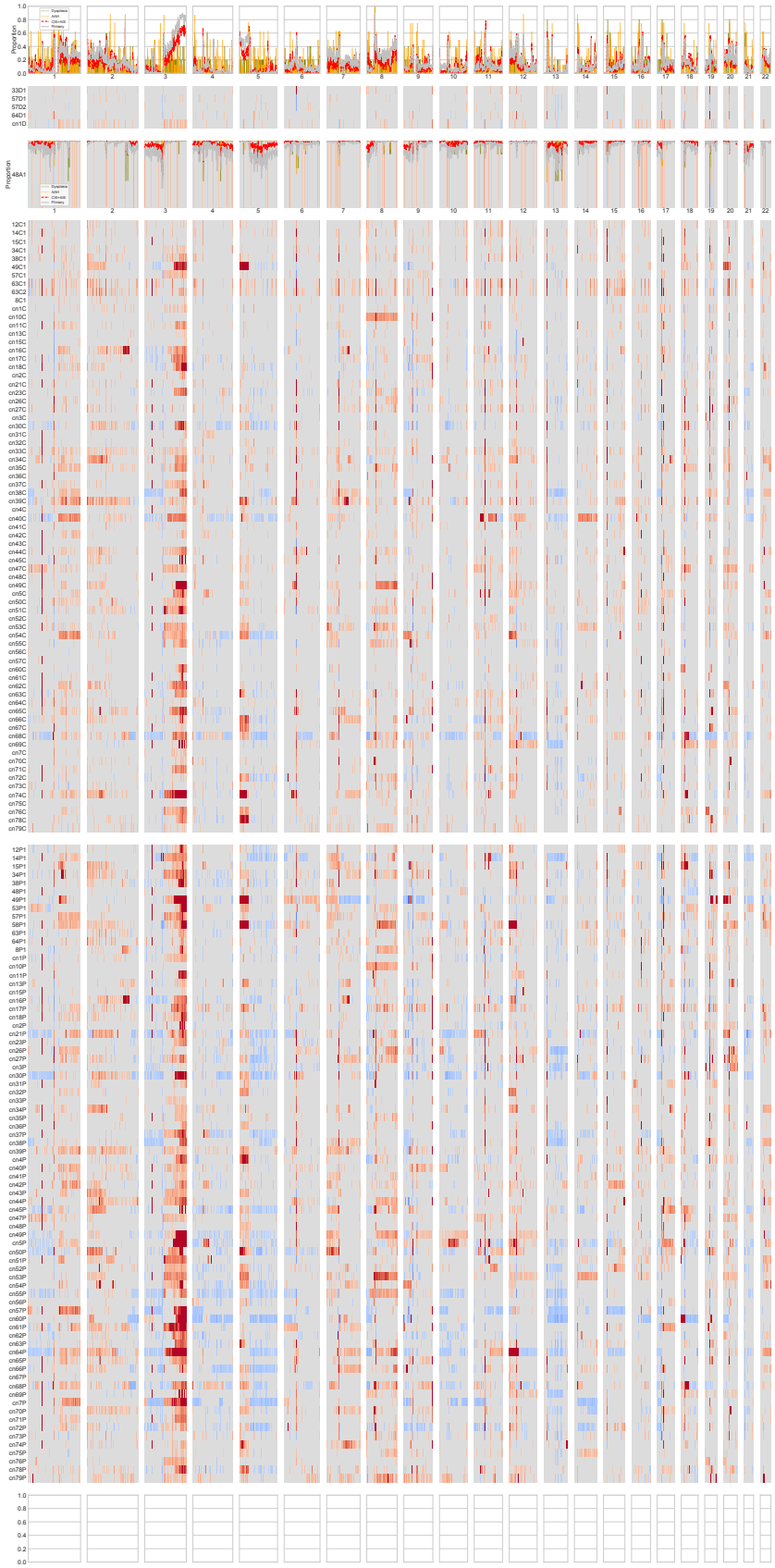


Figure 9: CNVkit LUSC Genome View Plot

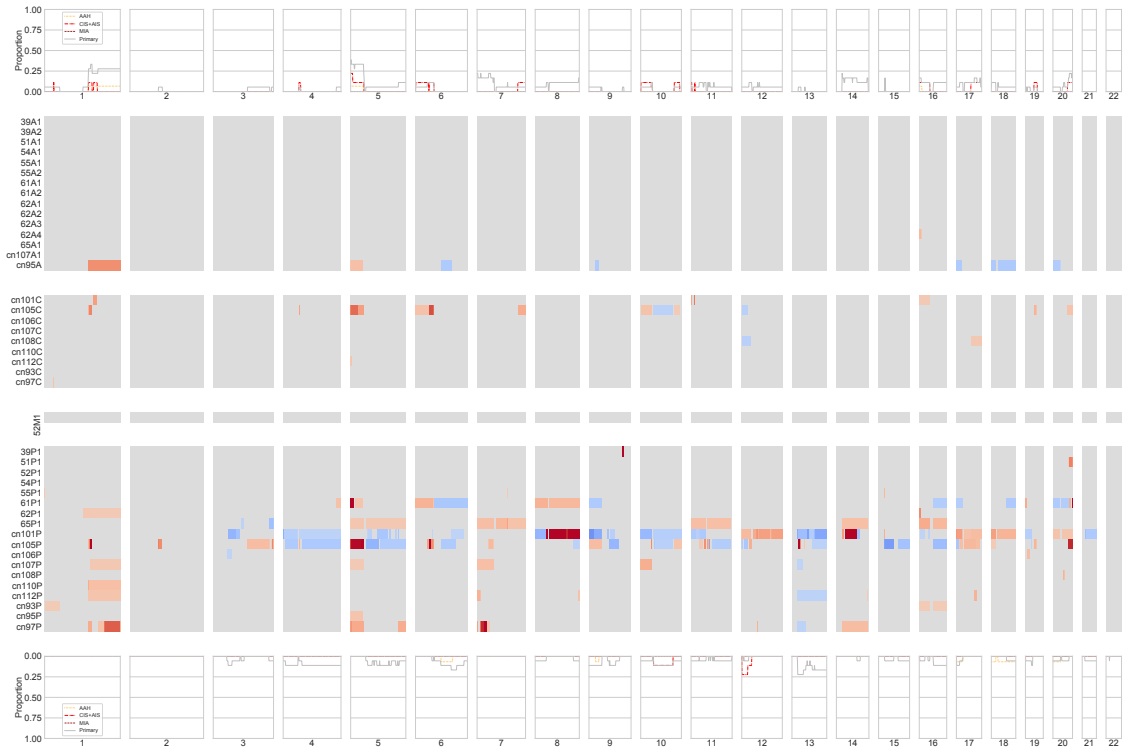


Figure 10: Sequenza LUAD Genome View Plot

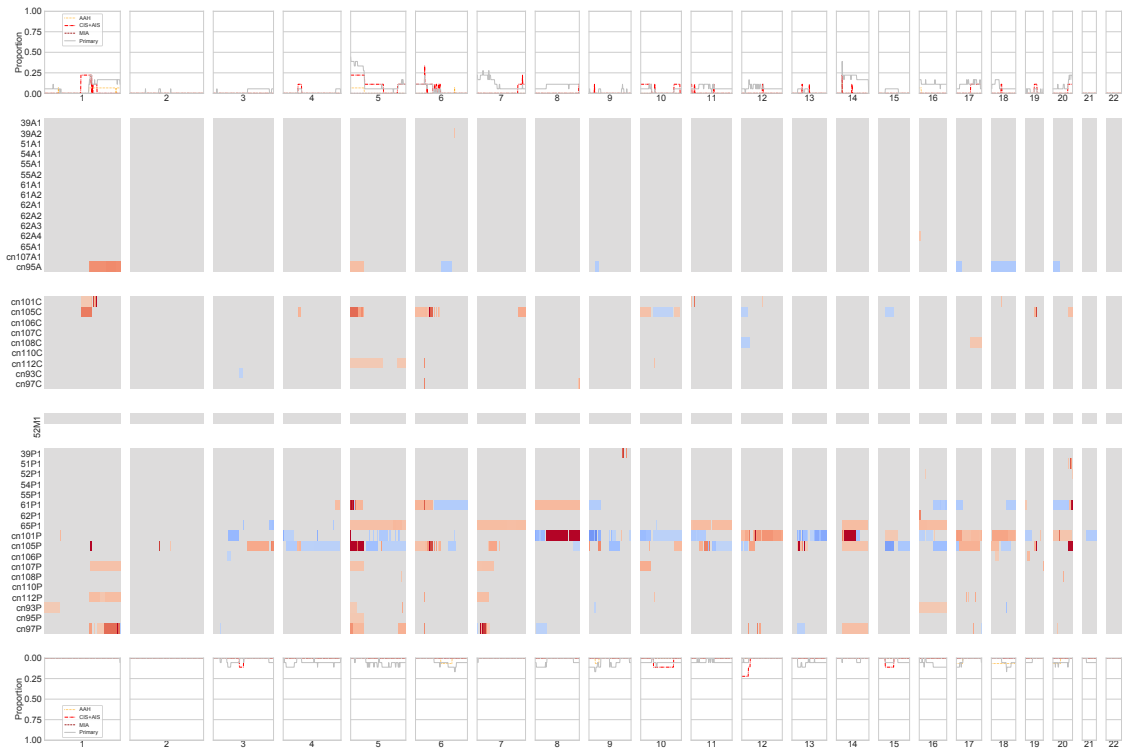


Figure 11: PureCN LUAD Genome View Plot

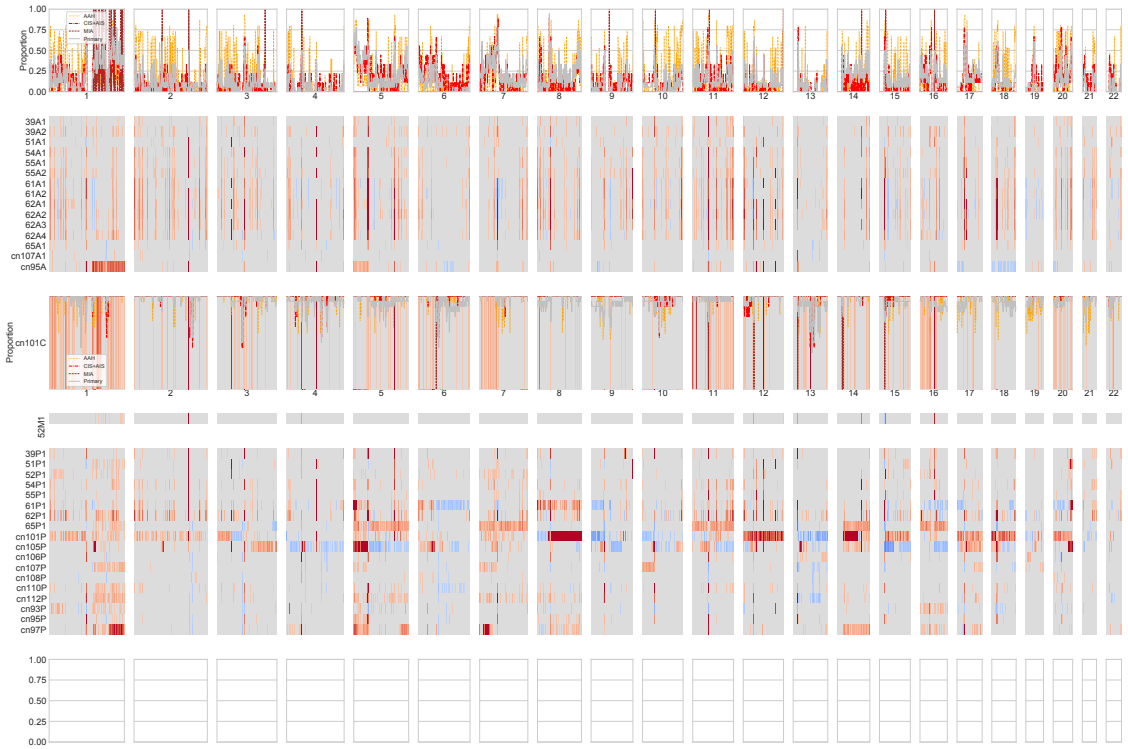


Figure 12: CNVkit LUAD Genome View Plot

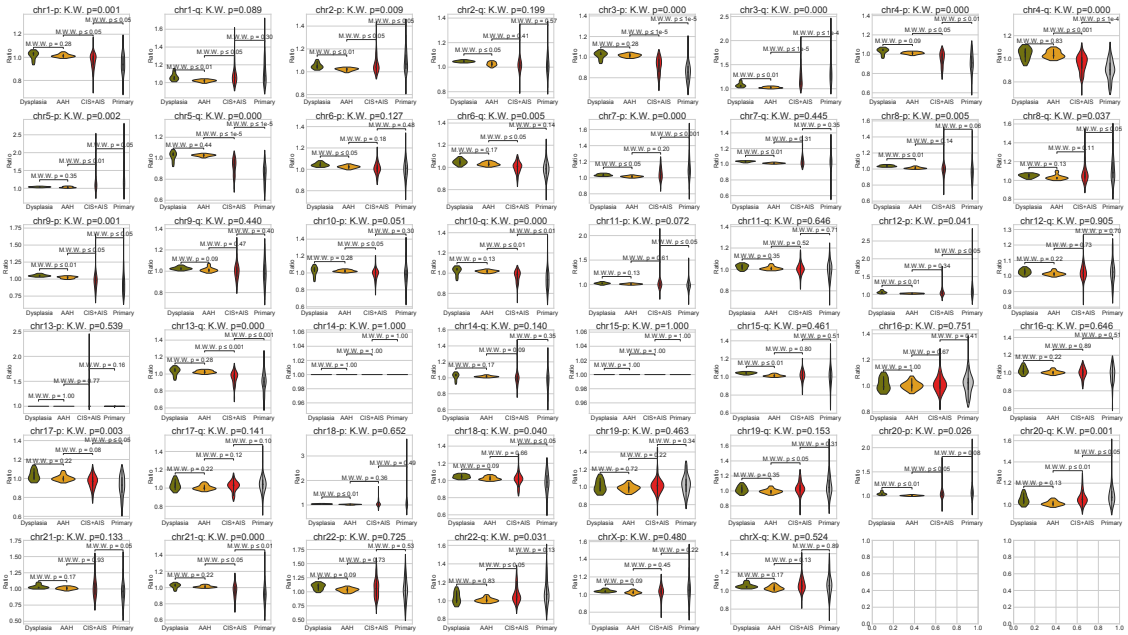


Figure 13: Sequenza LUSC Violin Plots

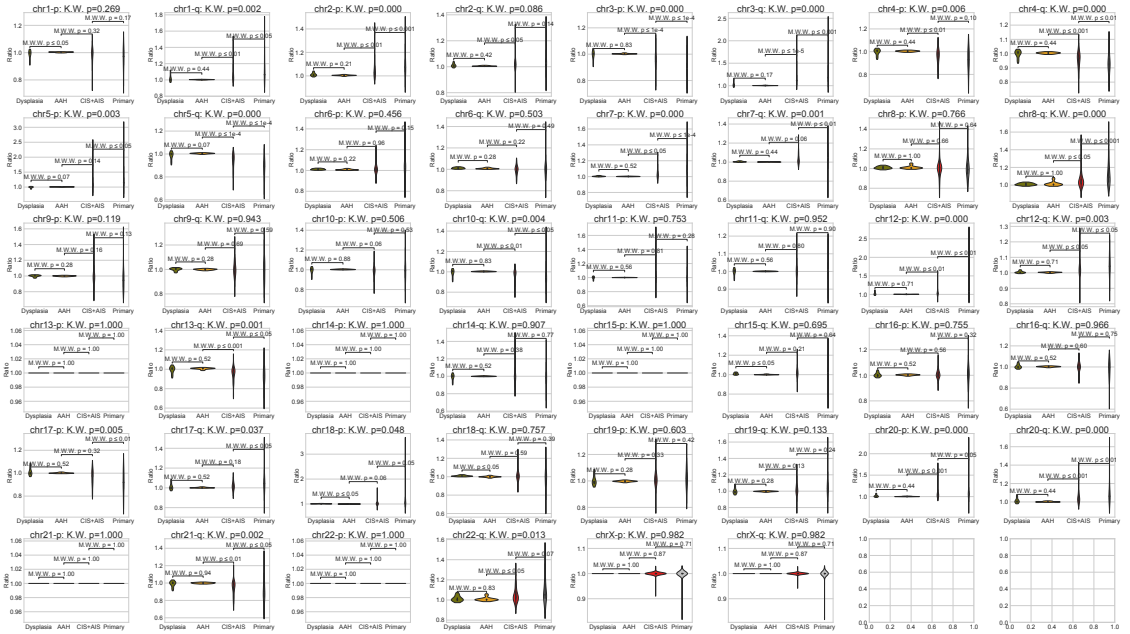


Figure 14: PureCN LUSC Violin Plots

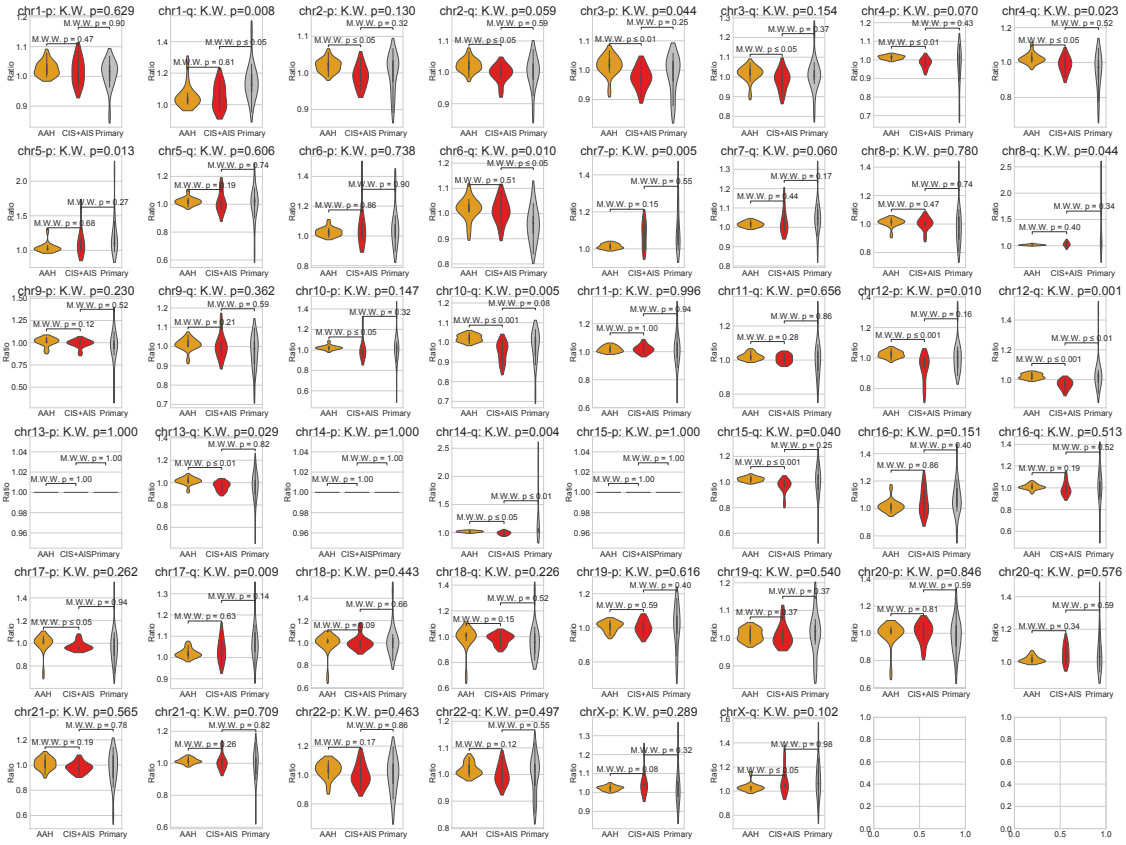


Figure 15: Sequenza LUAD Violin Plots

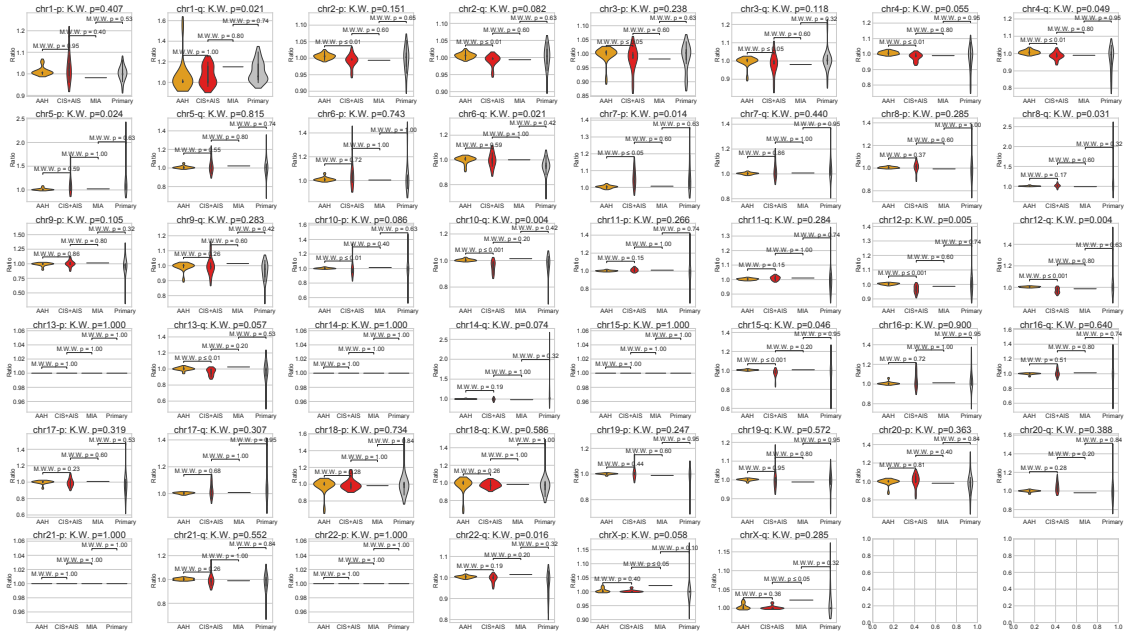
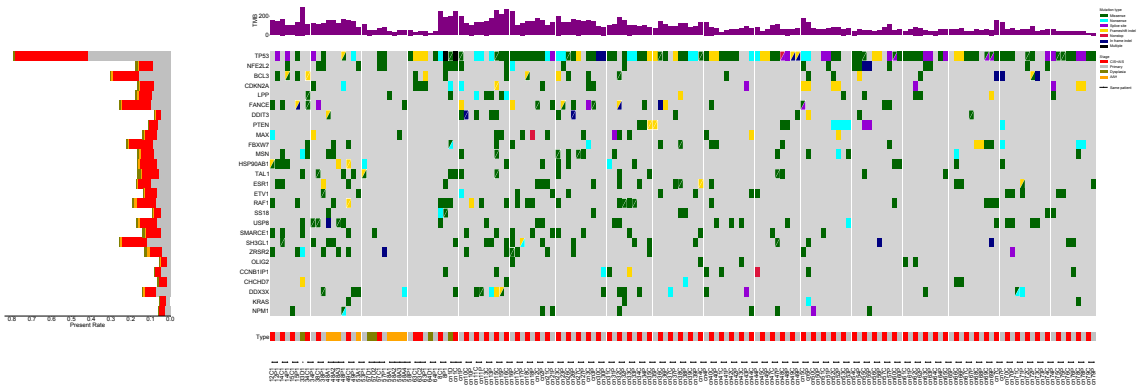


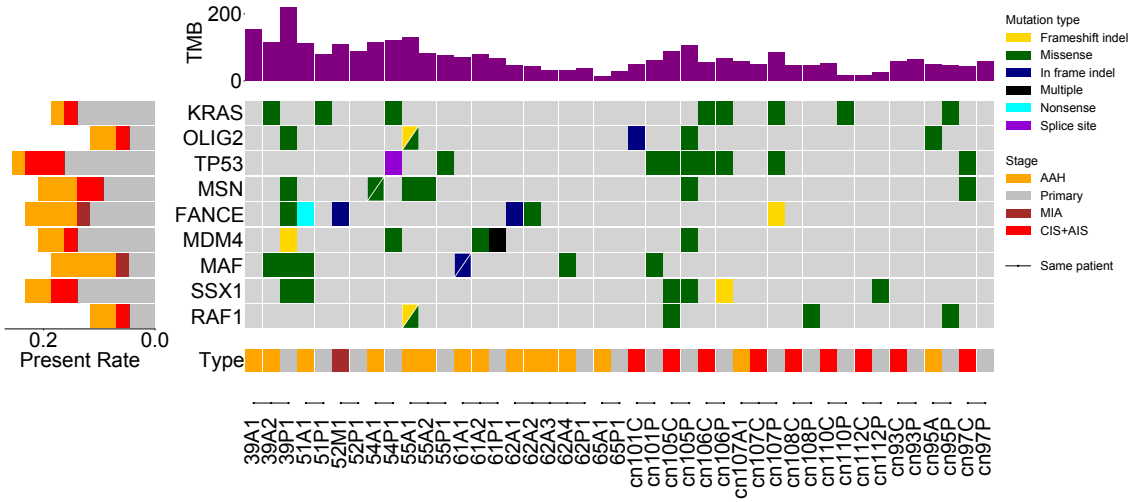
Figure 16: PureCN LUAD Violin Plots



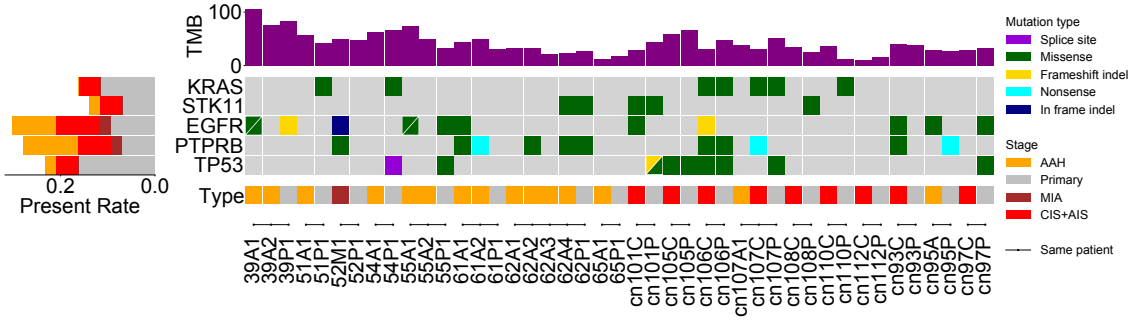
(a) BWA

(b) Bowtie2

Figure 17: Comut Plot by LUSC

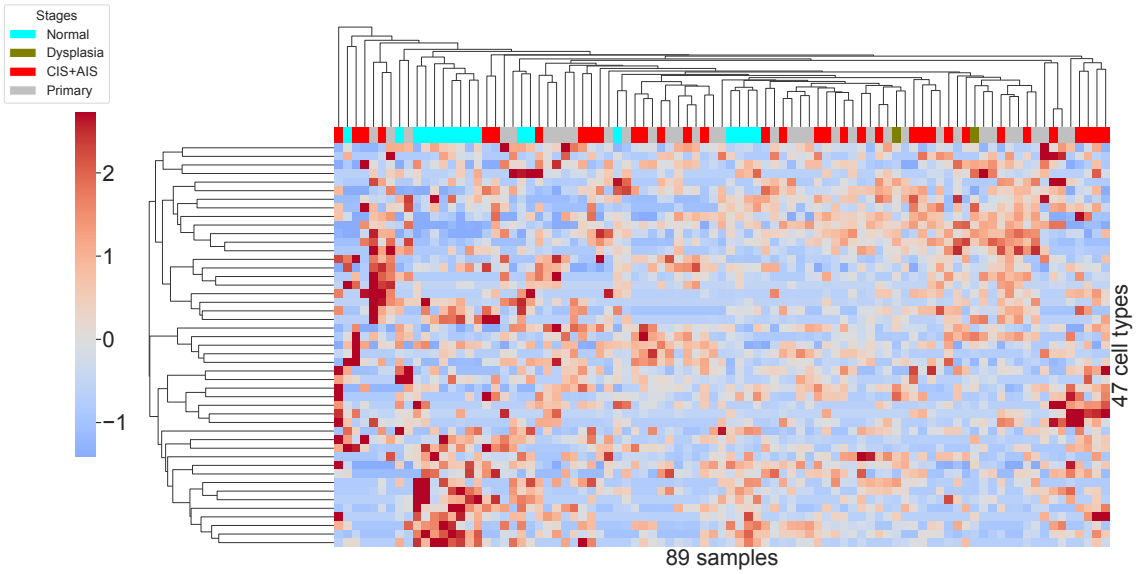


(a) BWA

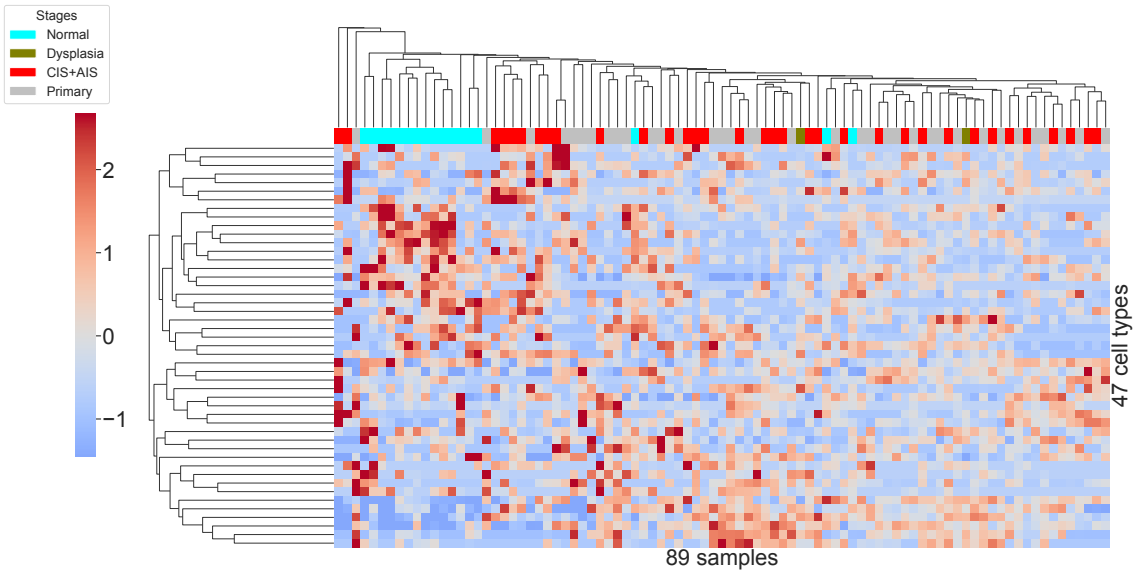


(b) Bowtie2

Figure 18: Comut Plot by LUAD

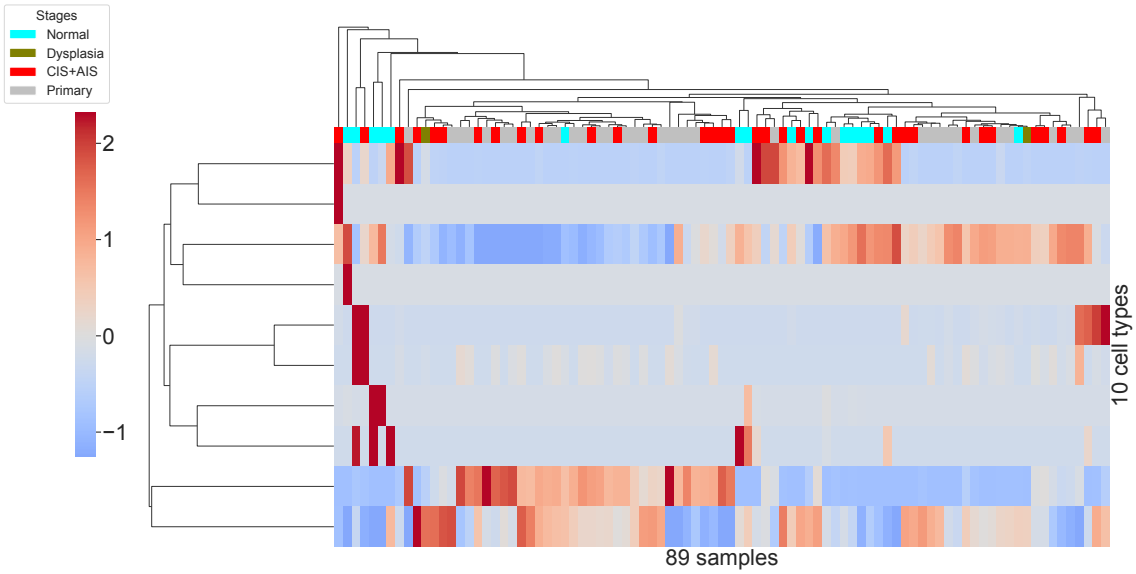


(a) STAR

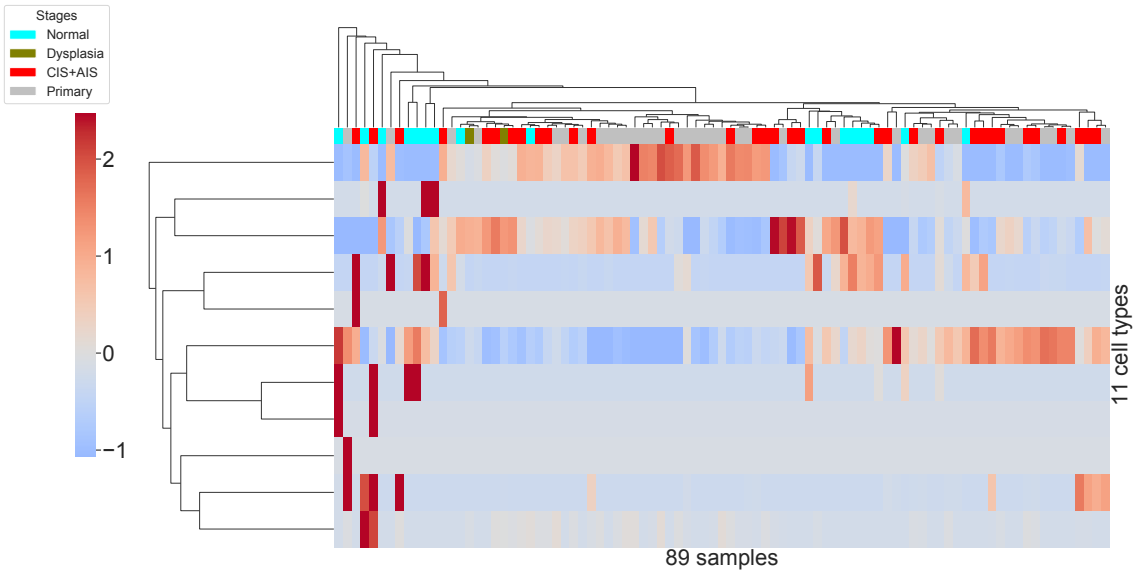


(b) Bowtie2

Figure 19: BisqueRNA clustermap plot with LUSC samples upon GSE131907

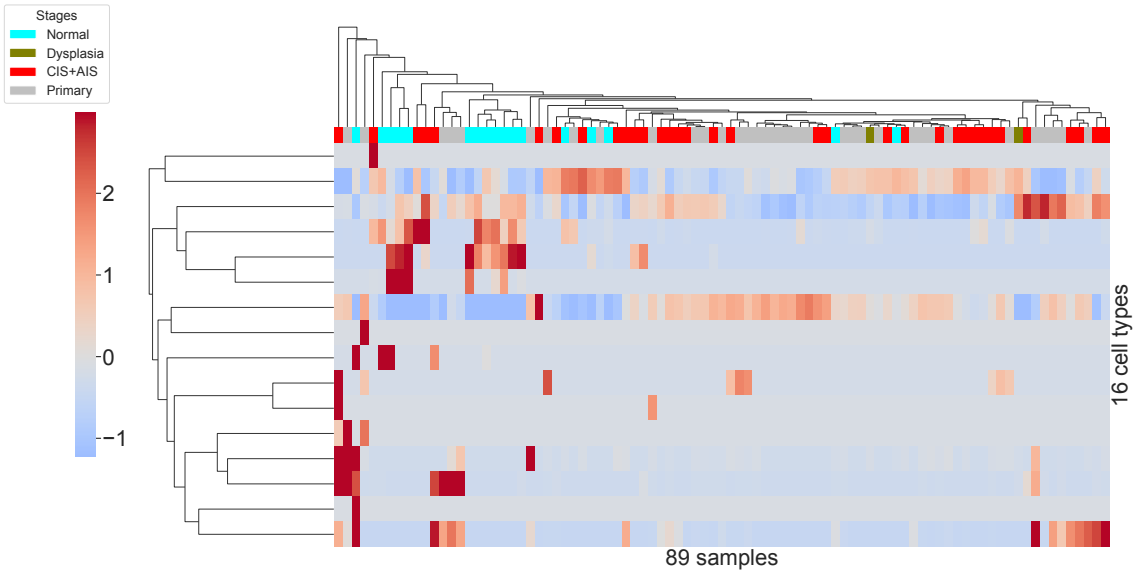


(a) STAR

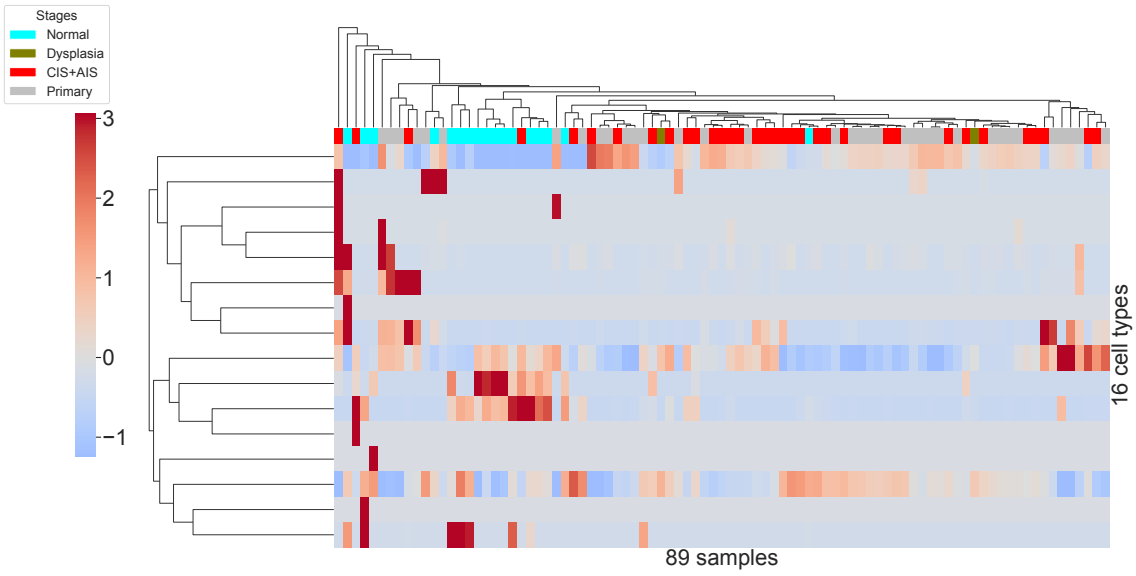


(b) Bowtie2

Figure 20: MuSiC clustermap plot with LUSC samples upon GSE131907



(a) STAR

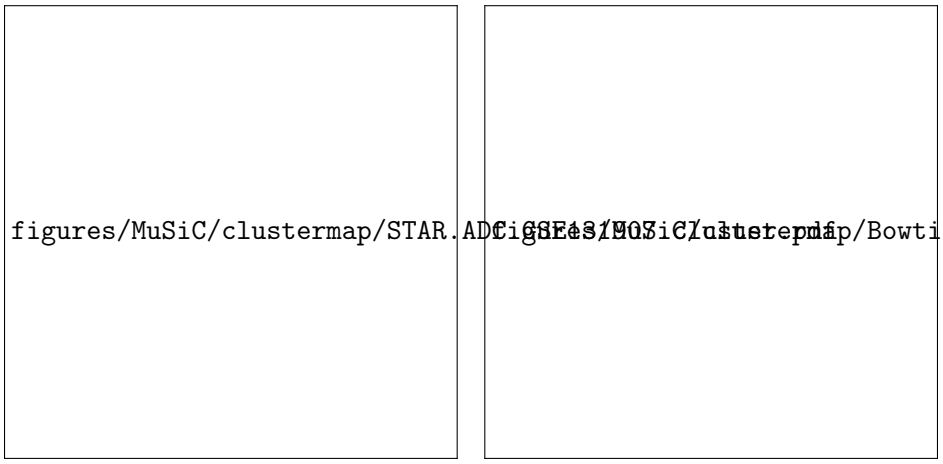


(b) Bowtie2

Figure 21: SCDC clustermap plot with LUSC samples upon GSE131907

(a) STAR (b) Bowtie2

Figure 22: BisqueRNA clustermap plot with LUAD samples upon GSE131907

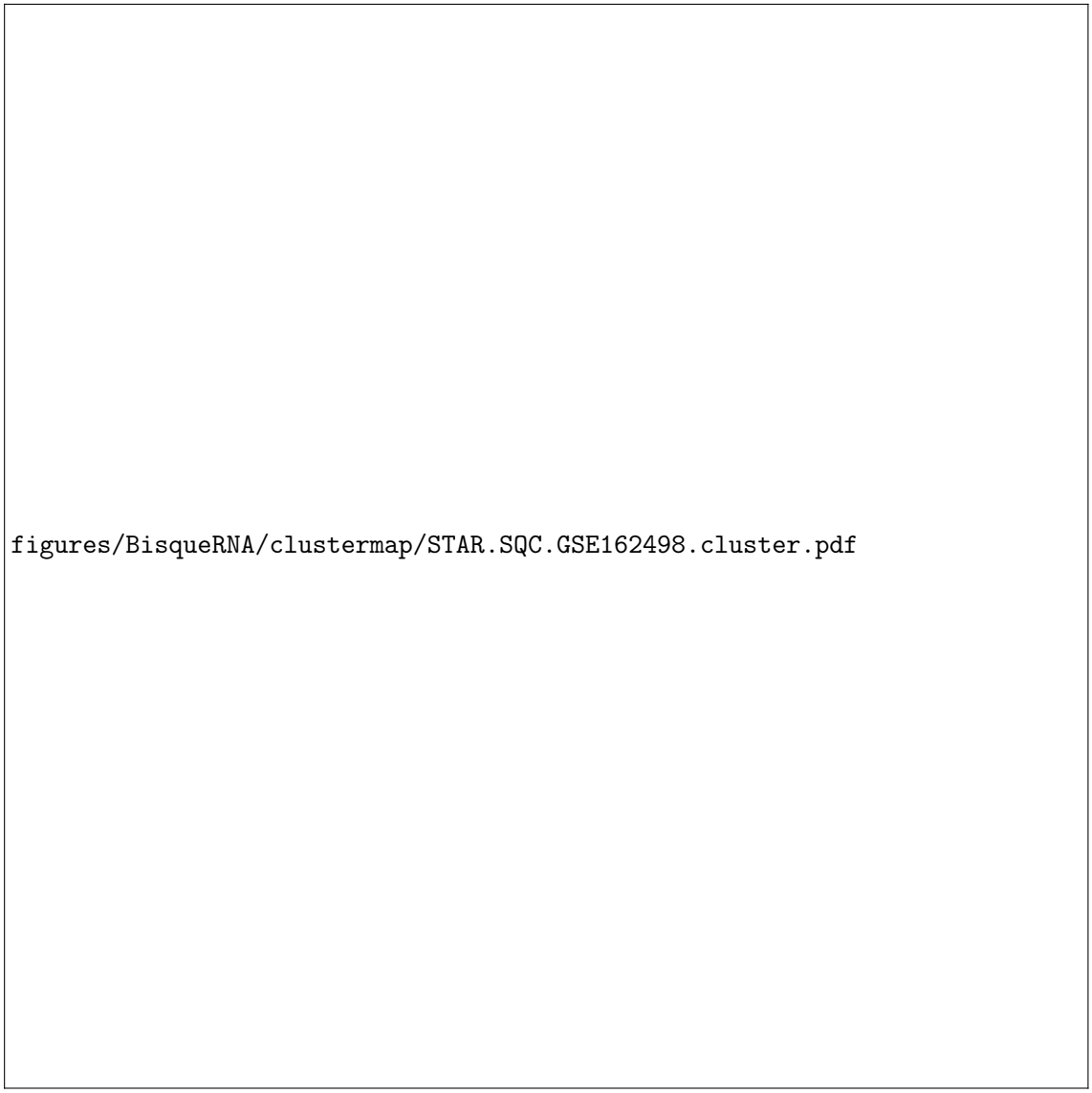


(a) STAR (b) Bowtie2

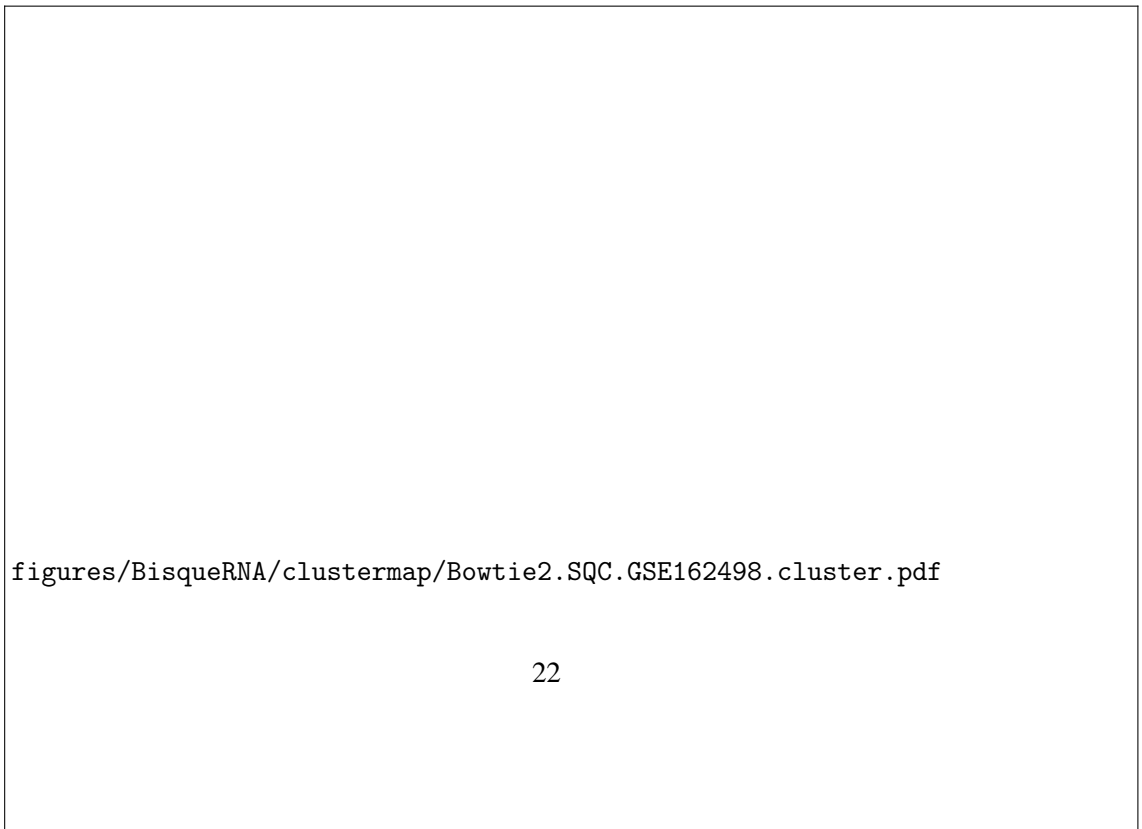
Figure 23: MuSiC clustermap plot with LUAD samples upon GSE131907

(a) STAR (b) Bowtie2

Figure 24: SCDCSiC clustermap plot with LUAD samples upon GSE131907

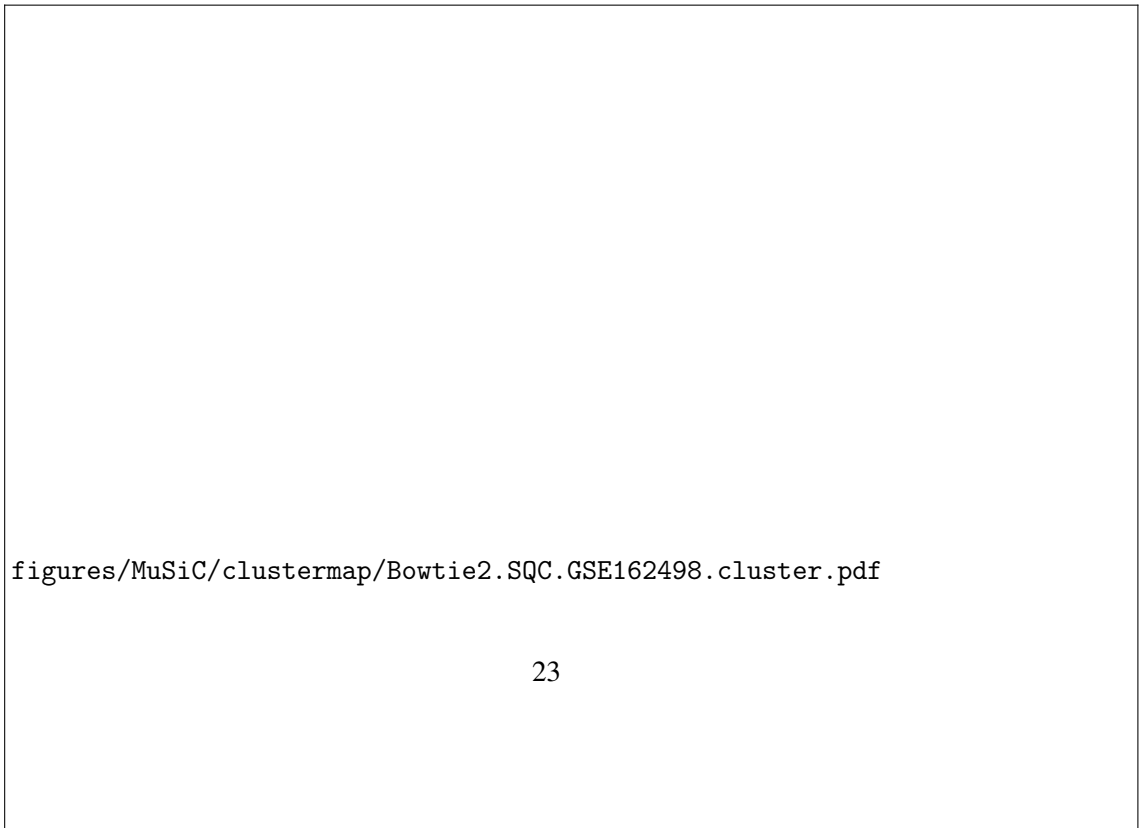

figures/BisqueRNA/clustermap/STAR.SQC.GSE162498.cluster.pdf

(a) STAR


figures/BisqueRNA/clustermap/Bowtie2.SQC.GSE162498.cluster.pdf


figures/MuSiC/clustermap/STAR.SQC.GSE162498.cluster.pdf

(a) STAR


figures/MuSiC/clustermap/Bowtie2.SQC.GSE162498.cluster.pdf

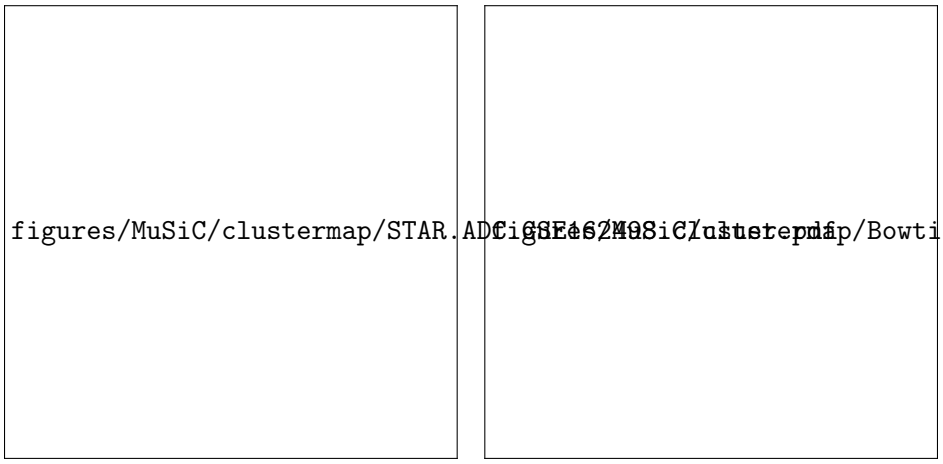

figures/SCDC/clustermap/STAR.SQC.GSE162498.cluster.pdf

(a) STAR


figures/SCDC/clustermap/Bowtie2.SQC.GSE162498.cluster.pdf

(a) STAR (b) Bowtie2

Figure 28: BisqueRNA clustermap plot with LUAD samples upon GSE162498



(a) STAR (b) Bowtie2

Figure 29: MuSiC clustermap plot with LUAD samples upon GSE162498

(a) STAR (b) Bowtie2

Figure 30: SCDC clustermap plot with LUAD samples upon GSE162498

V Discussion

5.1 General Conclusions

5.2 Plan for Future

5.3 Future Perspective

References

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.

Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., ... Rosell, R. (2015). Non-small-cell lung cancer. *Nature reviews Disease primers*, *1*(1), 1–16.

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.