

# Lung Pre-cancer

Jaewoong Lee

Sabin Park

Yeonsong Choi

Ilsun Yun

Semin Lee

2021-09-13

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Lung Cancer . . . . .	4
1.2	Precancer . . . . .	4
1.3	Study Objectives . . . . .	4
<b>2</b>	<b>Materials</b>	<b>4</b>
2.1	List of IPNs . . . . .	4
2.1.1	Carcinoma <i>in situ</i> . . . . .	4
2.1.2	Adenocarcinoma <i>in situ</i> . . . . .	4
2.1.3	Atypical Adenomatous Hyperplasia . . . . .	4
2.1.4	Dysplasia . . . . .	4
2.1.5	Minimally Invasive Adenocarcinoma . . . . .	4
2.2	Data Structure & Count . . . . .	10
<b>3</b>	<b>Methods</b>	<b>10</b>
3.1	Workflows . . . . .	10
<b>4</b>	<b>Results</b>	<b>10</b>
4.1	Quality Check . . . . .	10
4.2	Quality Check with FastQC . . . . .	10
4.2.1	Findings in Quality Check . . . . .	10
4.3	Copy Number Variations . . . . .	10
4.3.1	Copy Number Variation Analysis with Sequenza . . . . .	10
4.3.2	Cellularities and Ploidies . . . . .	10
4.3.3	Copy Number Variations . . . . .	10
4.3.4	Findings in Copy Number Variation Analysis . . . . .	25
4.4	Somatic Short Variation . . . . .	25
4.4.1	Somatic Short Variation Analysis with Mutect2 . . . . .	25
4.4.2	Findings in Somatic Short Variation Analysis . . . . .	25
4.5	Variant Allele Frequencies . . . . .	25
4.6	Differences in Gene Expression levels . . . . .	25
4.7	Bulk Cell Deconvolution . . . . .	25
4.7.1	Single-cell Reference Data . . . . .	25
4.7.2	CIBERSORTx . . . . .	25
4.7.3	BisqueRNA . . . . .	25
4.7.4	MuSiC . . . . .	25
4.7.5	SCDC . . . . .	25
<b>5</b>	<b>Discussion</b>	<b>25</b>
<b>6</b>	<b>References</b>	<b>25</b>

## List of Tables

## List of Figures

1	Common cancer survival rates (Hong et al., 2021) . . . . .	4
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	Example of FastQC Result (Andrews et al., 2012) . . . . .	6
7	FastQC results with WES data . . . . .	7
8	FastQC results with WTS data . . . . .	7
9	Representative Output of the Sequenza (Favero et al., 2015) . . . . .	7
10	Cellularities and Ploidies by BWA in ADC . . . . .	8
11	Cellularities and Ploidies by BWA in SQC . . . . .	8
12	Cellularities and Ploidies by Bowtie2 in ADC . . . . .	9

13	Cellularities and Ploidies by Bowtie2 in SQC . . . . .	9
14	CNV plot by BWA in ADC . . . . .	10
15	CNV plot by BWA in SQC . . . . .	11
16	CNV plot by Bowtie2 in ADC . . . . .	11
17	CNV plot by Bowtie2 in SQC . . . . .	12
18	Simple CNV plot by BWA in ADC . . . . .	12
19	Simple CNV plot by BWA in SQC . . . . .	12
20	Simple CNV plot by Bowtie2 in ADC . . . . .	12
21	Simple CNV plot by Bowtie2 in SQC . . . . .	13
22	Somatic Short Variant Discovery Workflow (Van der Auwera et al., 2013; DePristo et al., 2011) . . . . .	13
23	CoMut plot by BWA in ADC . . . . .	13
24	CoMut plot by BWA in SQC . . . . .	14
25	CoMut plot by Bowtie2 in ADC . . . . .	14
26	CoMut plot by Bowtie2 <sup>c</sup> in SQC . . . . .	14
27	DEG volcano plots by Bowtie2 in ADC . . . . .	15
28	DEG volcano plots by Bowtie2 in SQC . . . . .	15
29	DEG volcano plots by STAR in ADC . . . . .	16
30	DEG volcano plots by Bowtie2 in SQC . . . . .	16
31	DEG Venn Diagram by Bowtie2 in ADC . . . . .	17
32	DEG Venn Diagram by Bowtie2 in SQC . . . . .	17
33	DEG Venn Diagram by STAR in ADC . . . . .	17
34	DEG Venn Diagram by STAR in SQC . . . . .	18
35	Comprehensive dissection and clustering of 208,506 single cells from LUAD patients (Kim et al., 2020) . . . . .	18
36	Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in ADC . . . . .	19
37	Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in SQC . . . . .	19
38	Cell deconvolution clustermap by STAR and CIBERSORTx in ADC . . . . .	20
39	Cell deconvolution clustermap by STAR and CIBERSORTx in SQC . . . . .	20
40	Cell deconvolution clustermap by Bowtie2 and BisqueRNA in ADC . . . . .	21
41	Cell deconvolution clustermap by Bowtie2 and BisqueRNA in SQC . . . . .	21
42	Cell deconvolution clustermap by STAR and BisqueRNA in ADC . . . . .	22
43	Cell deconvolution clustermap by STAR and BisqueRNA in SQC . . . . .	22
44	Cell deconvolution clustermap by Bowtie2 and MuSiC in ADC . . . . .	23
45	Cell deconvolution clustermap by Bowtie2 and MuSiC in SQC . . . . .	23
46	Cell deconvolution clustermap by STAR and MuSiC in ADC . . . . .	24
47	Cell deconvolution clustermap by STAR and MuSiC in SQC . . . . .	24
48	Cell deconvolution clustermap by Bowtie2 and SCDC in ADC . . . . .	25
49	Cell deconvolution clustermap by Bowtie2 and SCDC in SQC . . . . .	26
50	Cell deconvolution clustermap by STAR and SCDC in ADC . . . . .	26
51	Cell deconvolution clustermap by STAR and SCDC in SQC . . . . .	27

# 1 Introduction

## 1.1 Lung Cancer

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

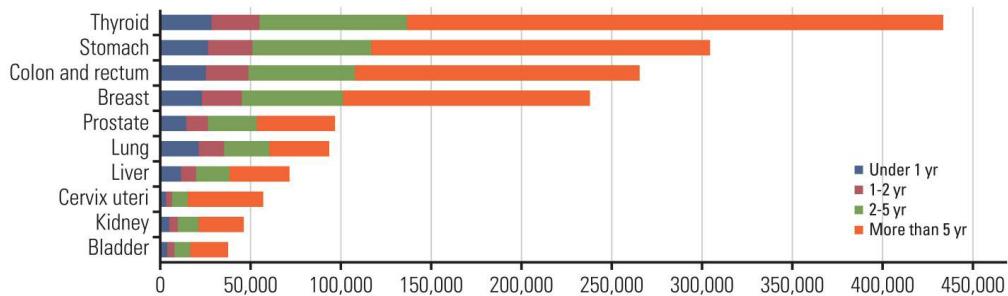


Figure 1: Common cancer survival rates (Hong et al., 2021)

## 1.2 Precancer

### 1.3 Study Objectives

## 2 Materials

### 2.1 List of IPNs

#### 2.1.1 Carcinoma *in situ*

Carcinoma *in situ* (CIS)

#### 2.1.2 Adenocarcinoma *in situ*

Adenocarcinoma *in situ* (AIS)

#### 2.1.3 Atypical Adenomatous Hyperplasia

Atypical adenomatous hyperplasia (AAH)

#### 2.1.4 Dysplasia

#### 2.1.5 Minimally Invasive Adenocarcinoma

Minimally invasive adenocarcinoma (MIA)

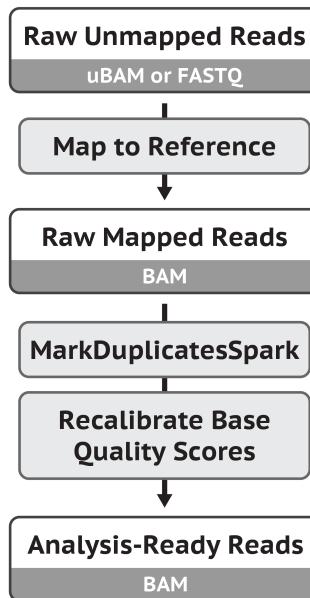


Figure 2: Workflow for data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

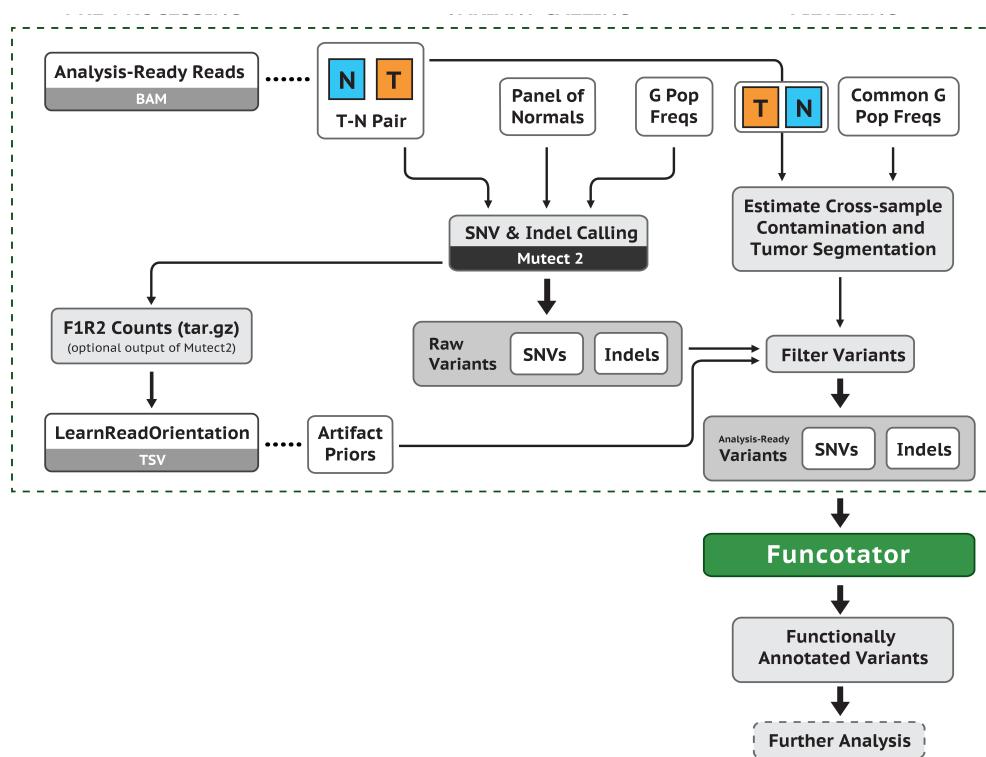


Figure 3: Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

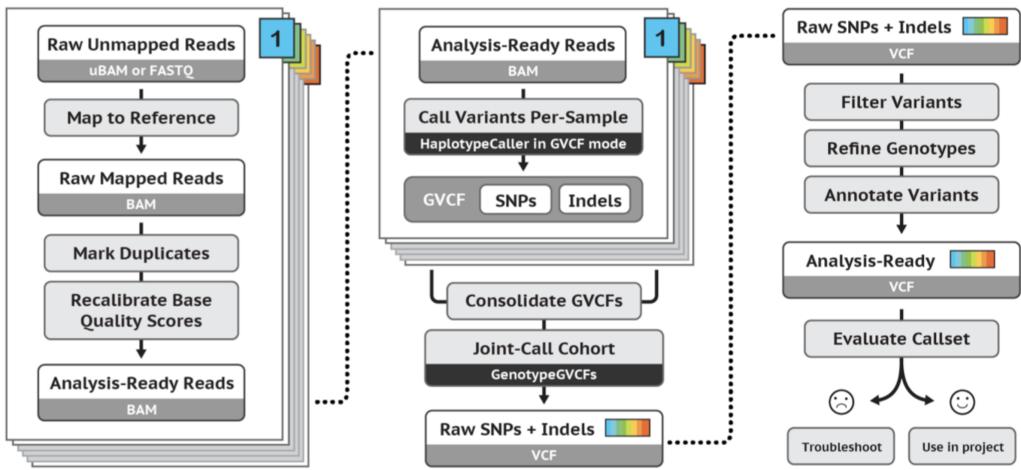


Figure 4: Germline short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

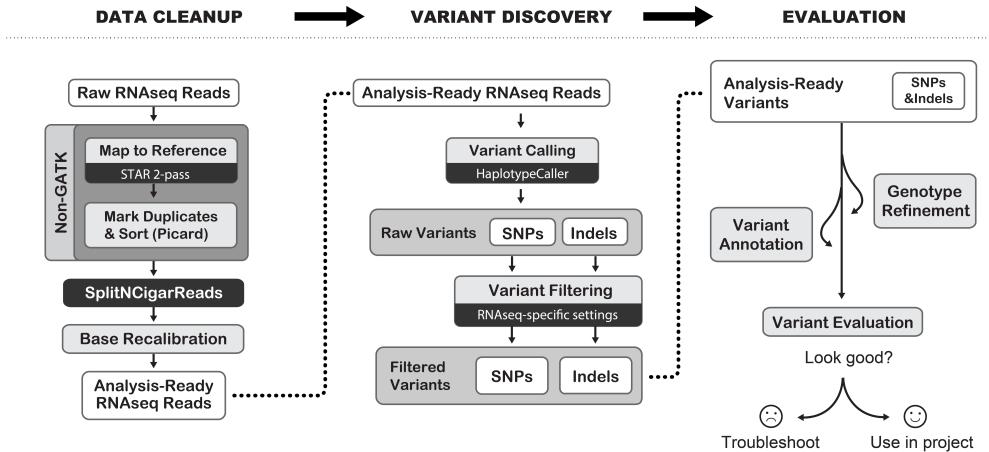


Figure 5: RNA-seq short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

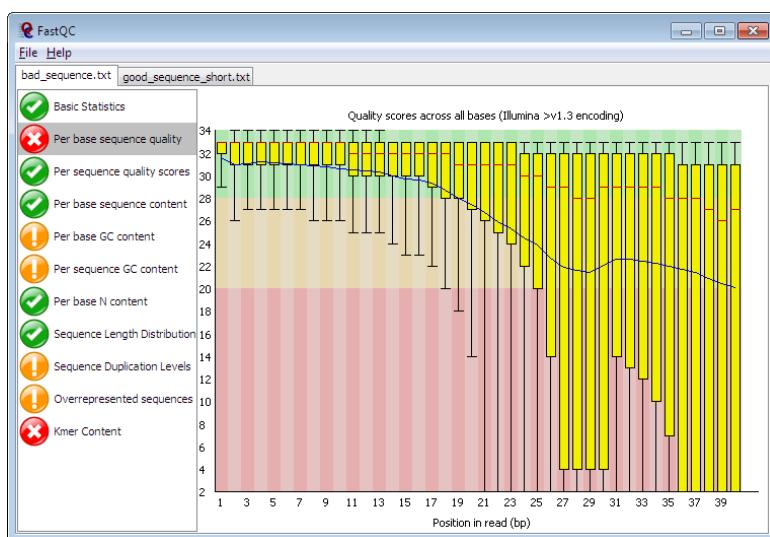


Figure 6: Example of FastQC Result (Andrews et al., 2012)

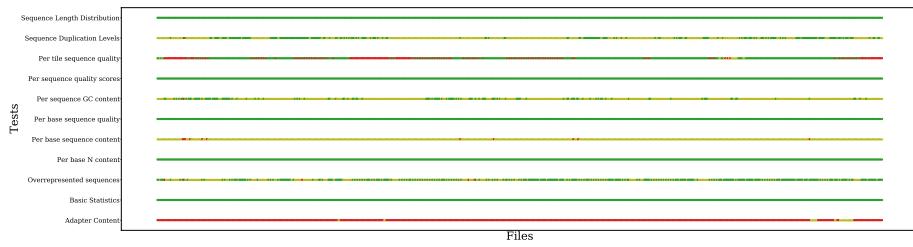


Figure 7: FastQC results with WES data

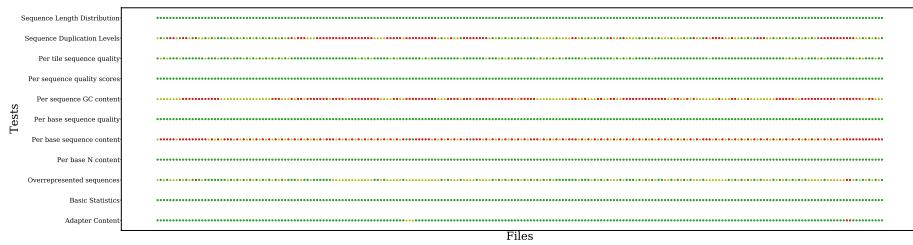


Figure 8: FastQC results with WTS data

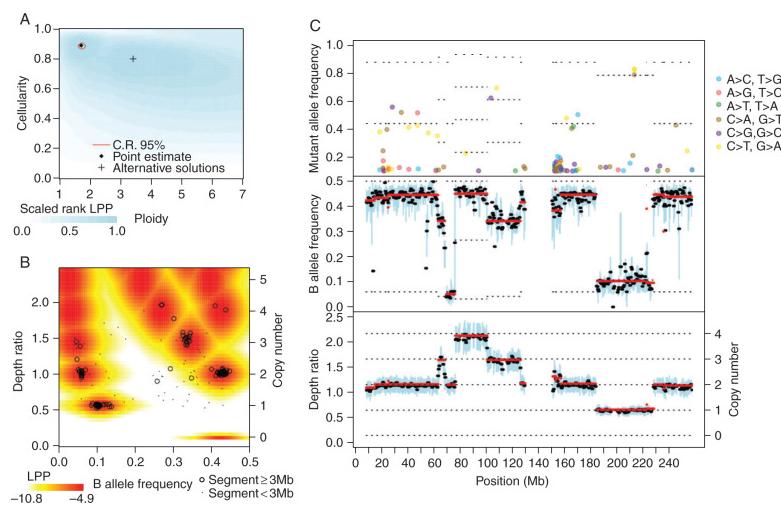


Figure 9: Representative Output of the Sequenza (Favero et al., 2015)

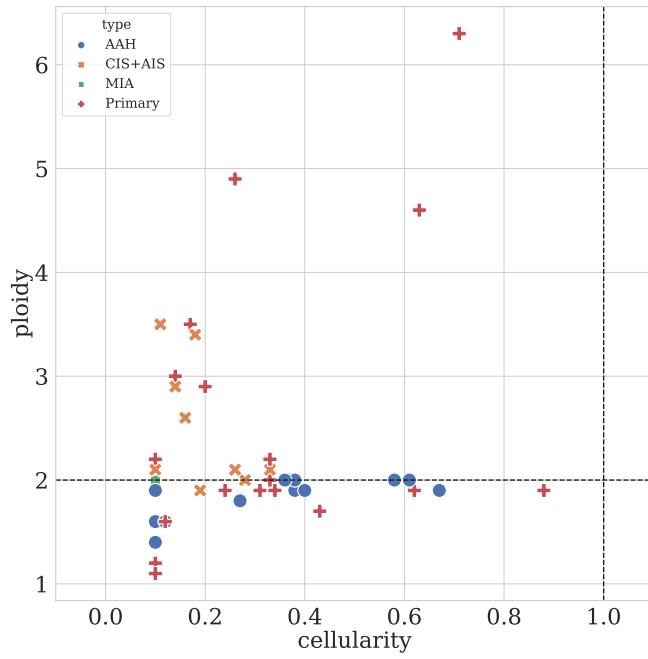


Figure 10: Cellularities and Ploidies by BWA in ADC

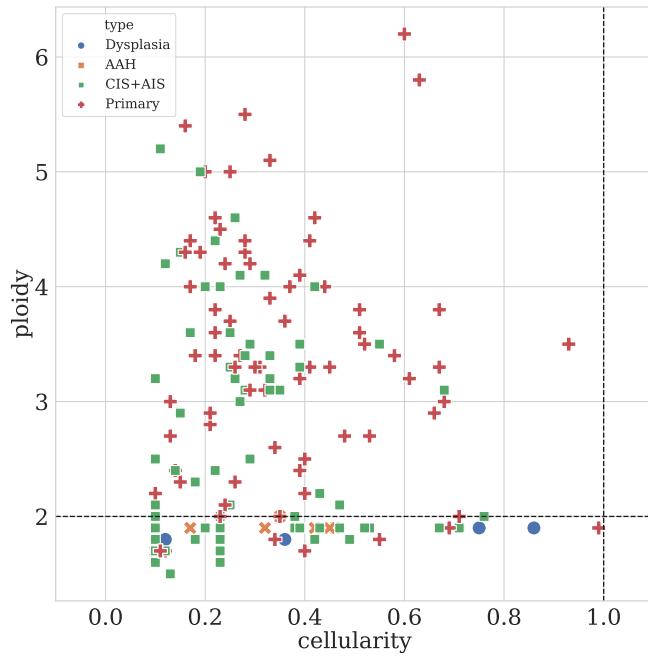


Figure 11: Cellularities and Ploidies by BWA in SQC

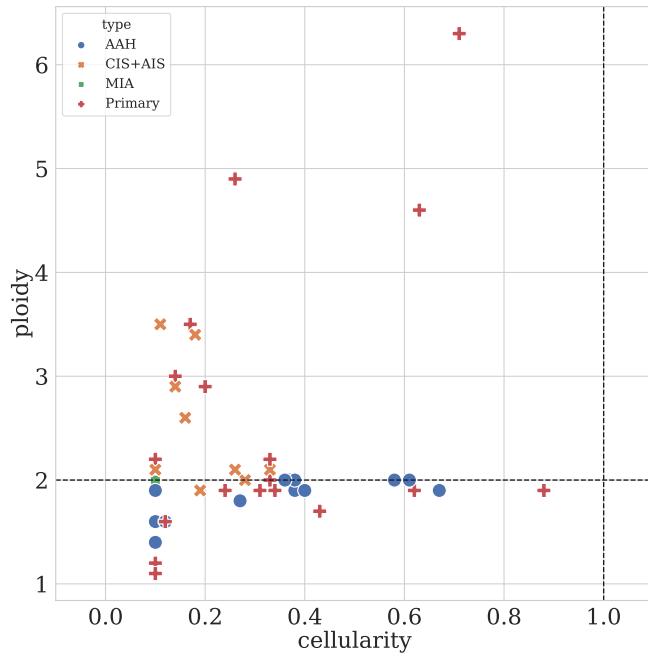


Figure 12: Cellularities and Ploidies by Bowtie2 in ADC

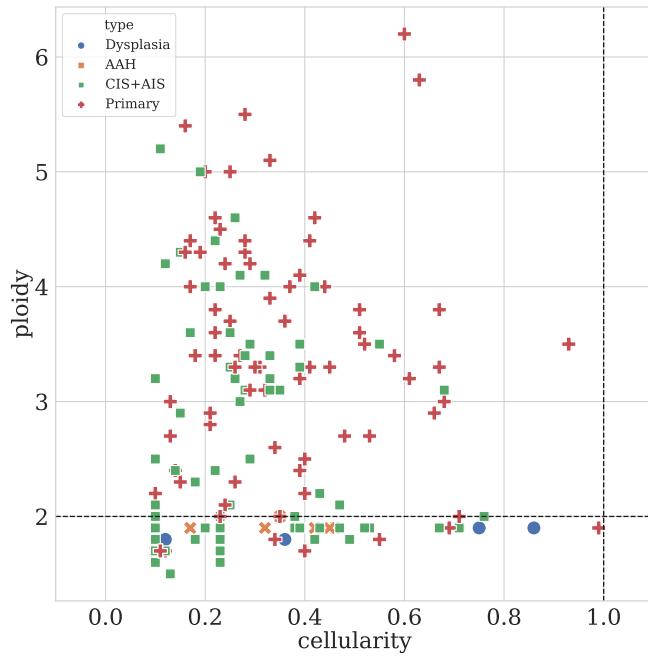


Figure 13: Cellularities and Ploidies by Bowtie2 in SQC

## 2.2 Data Structure & Count

# 3 Methods

## 3.1 Workflows

# 4 Results

## 4.1 Quality Check

## 4.2 Quality Check with FastQC

### 4.2.1 Findings in Quality Check

## 4.3 Copy Number Variations

### 4.3.1 Copy Number Variation Analysis with Sequenza

### 4.3.2 Cellularities and Ploidies

### 4.3.3 Copy Number Variations

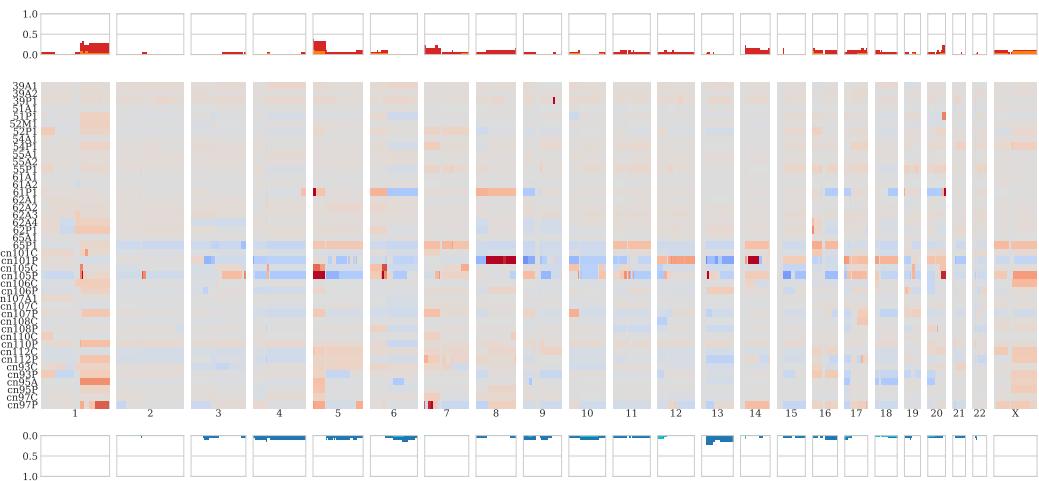


Figure 14: CNV plot by BWA in ADC

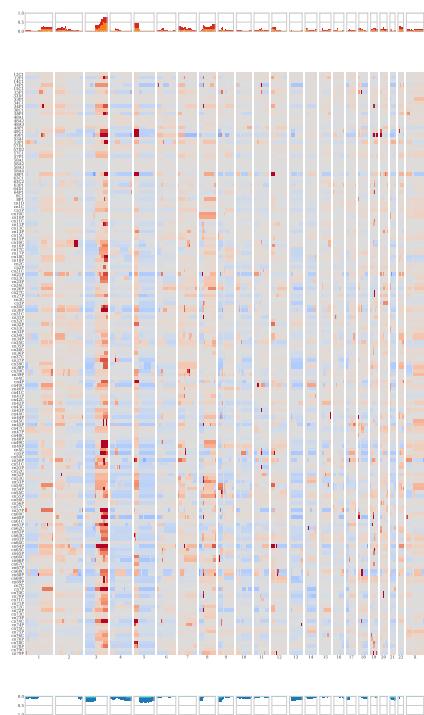


Figure 15: CNV plot by BWA in SQC

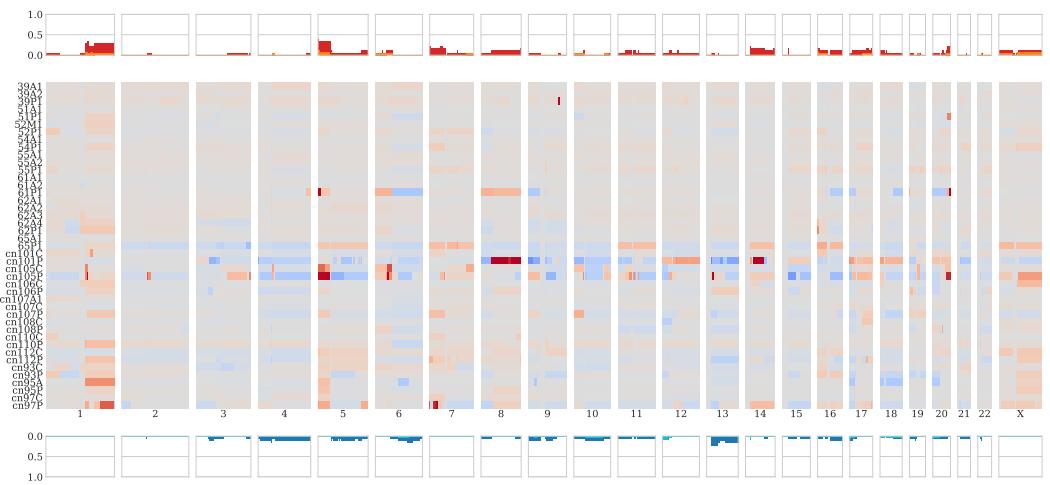


Figure 16: CNV plot by Bowtie2 in ADC

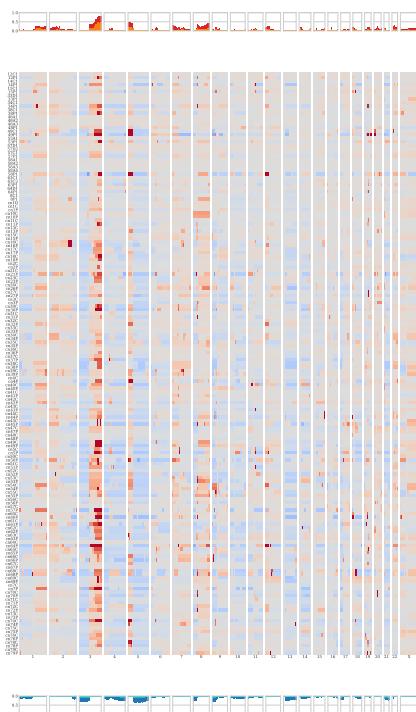


Figure 17: CNV plot by Bowtie2 in SQC

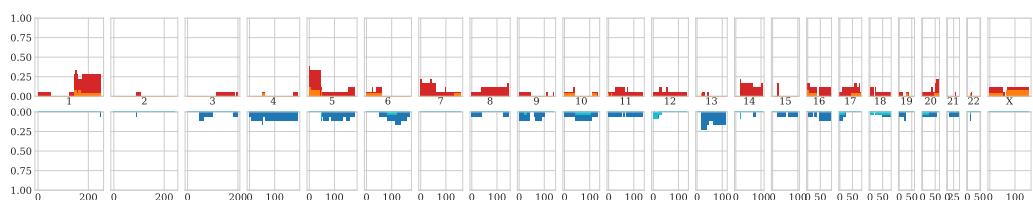


Figure 18: Simple CNV plot by BWA in ADC

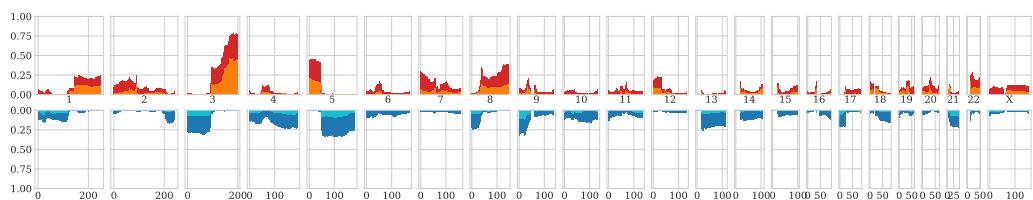


Figure 19: Simple CNV plot by BWA in SQC

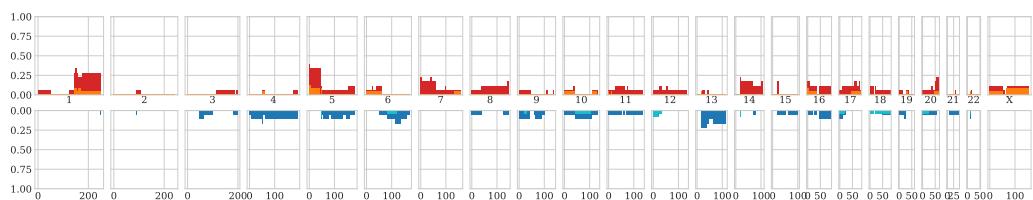


Figure 20: Simple CNV plot by Bowtie2 in ADC

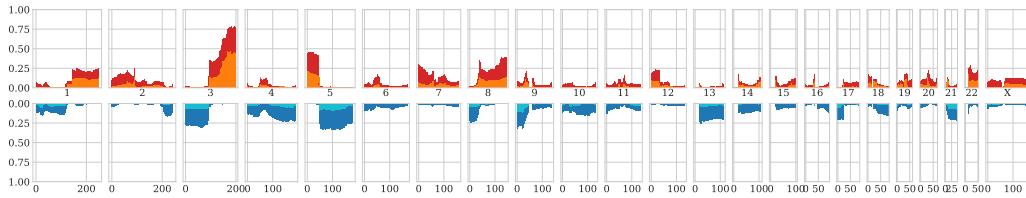


Figure 21: Simple CNV plot by Bowtie2 in SQC

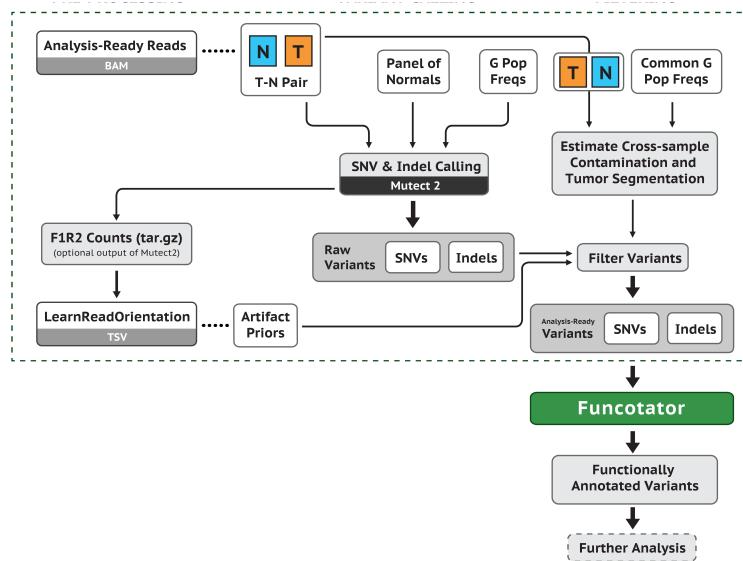


Figure 22: Somatic Short Variant Discovery Workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

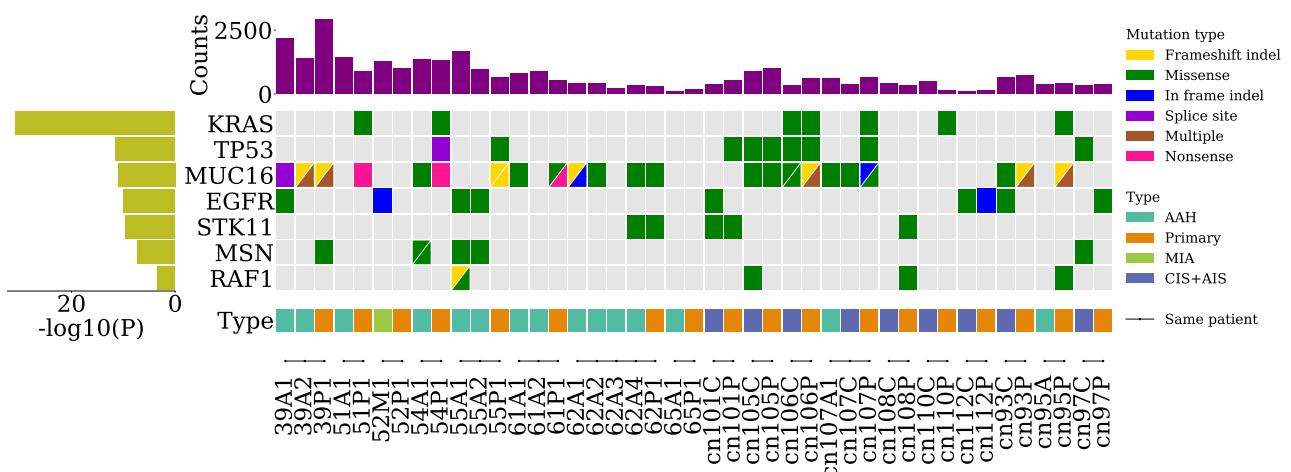


Figure 23: CoMut plot by BWA in ADC



Figure 24: CoMut plot by BWA in SQC

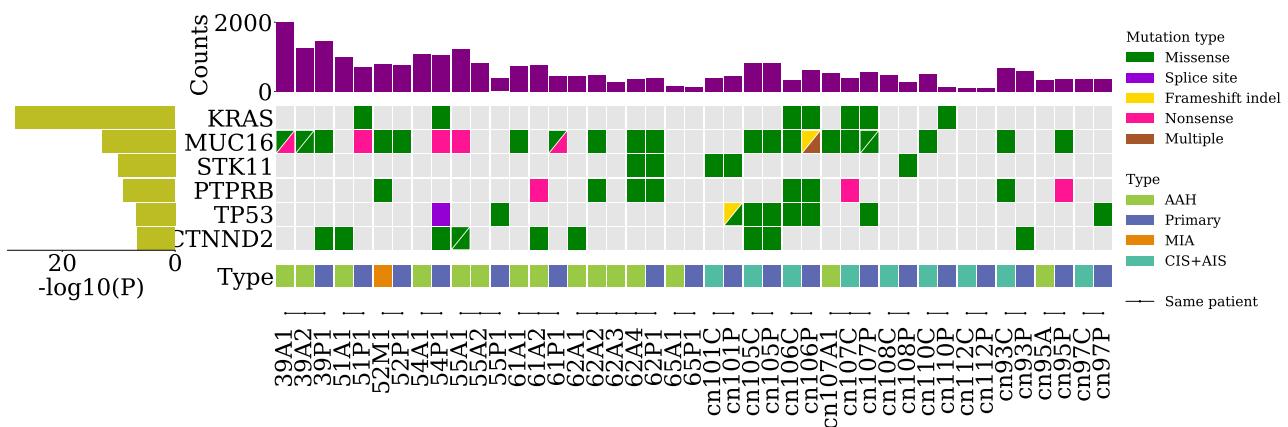


Figure 25: CoMut plot by Bowtie2 in ADC

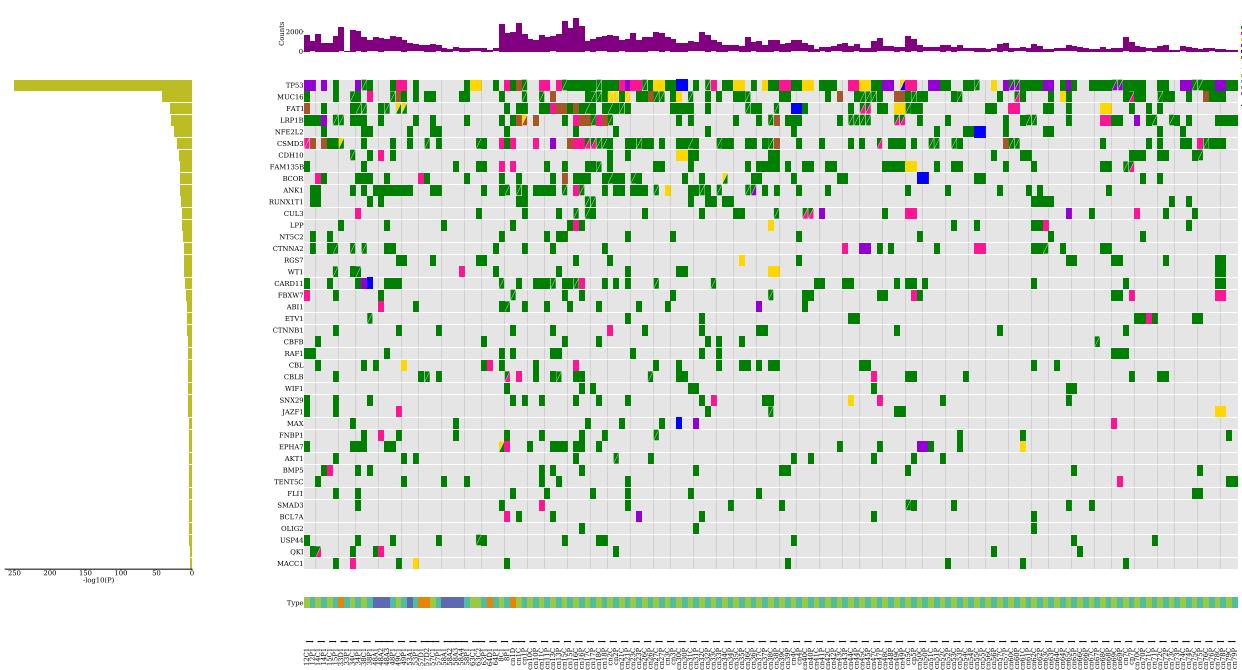


Figure 26: CoMut plot by Bowtie2\* in SQC

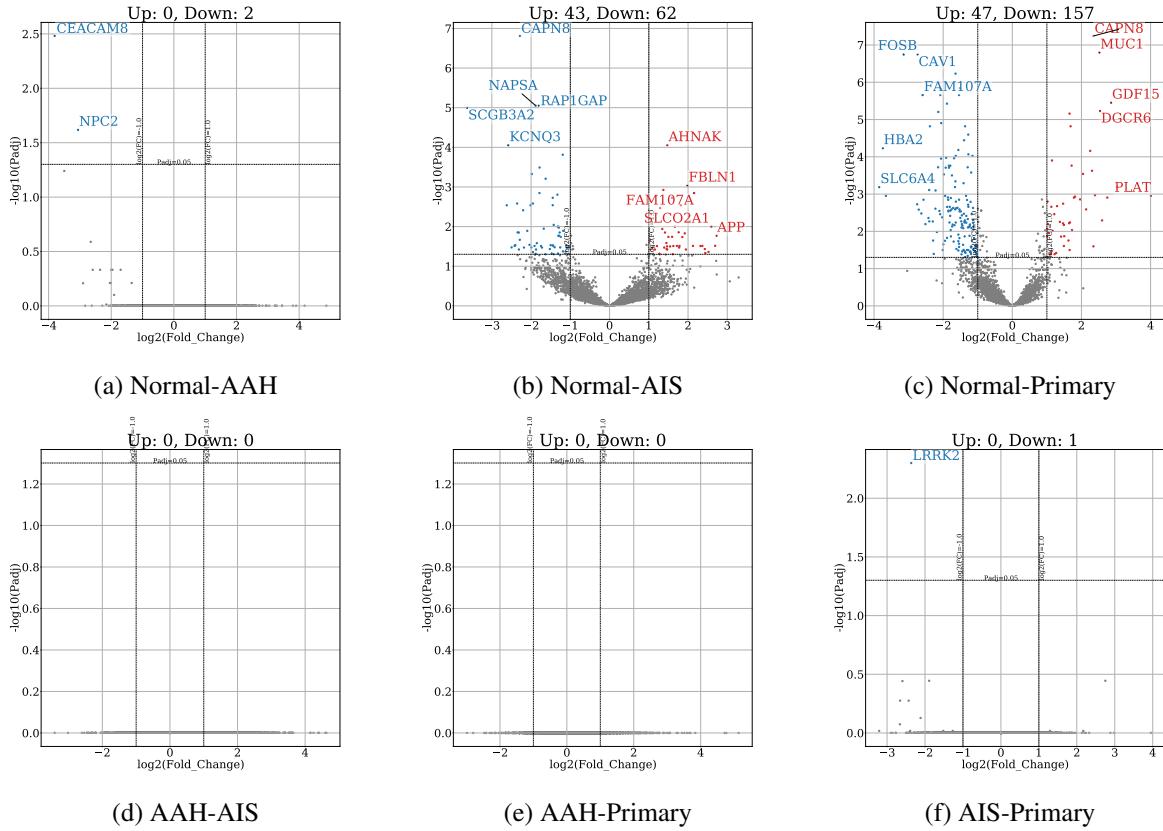


Figure 27: DEG volcano plots by Bowtie2 in ADC

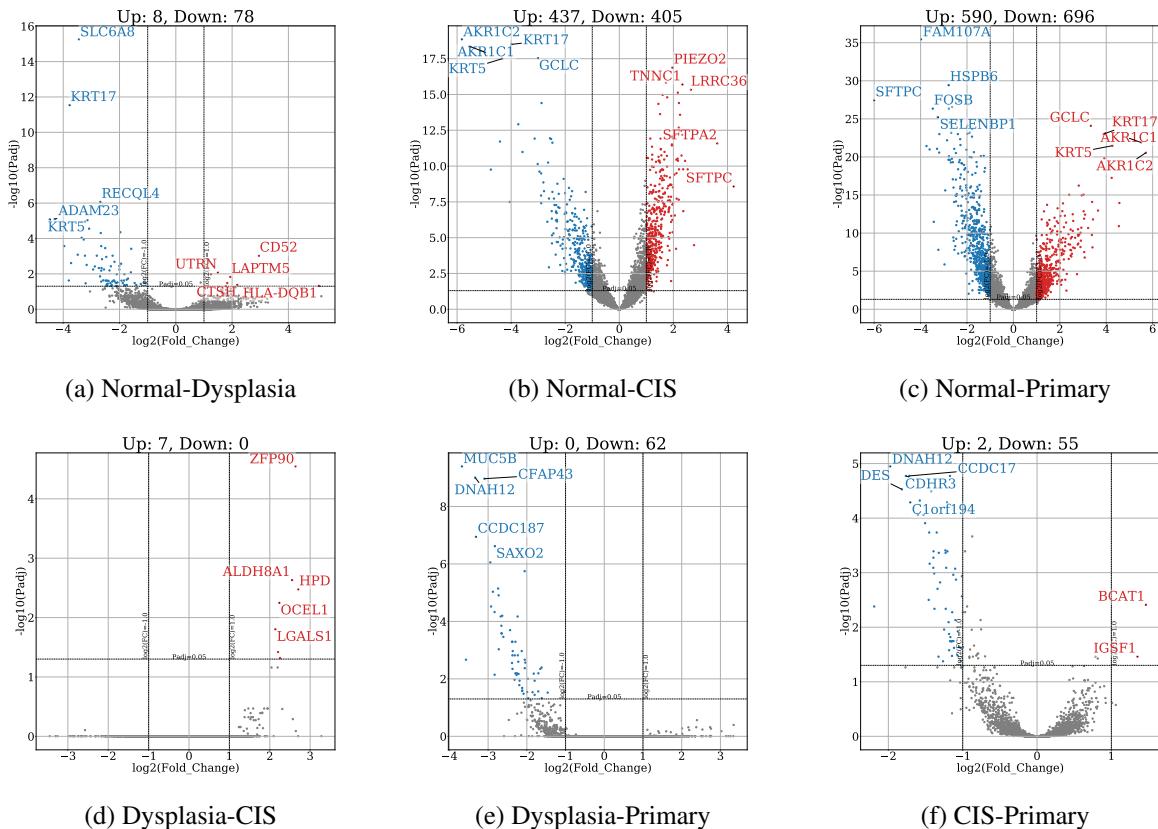


Figure 28: DEG volcano plots by Bowtie2 in SQC

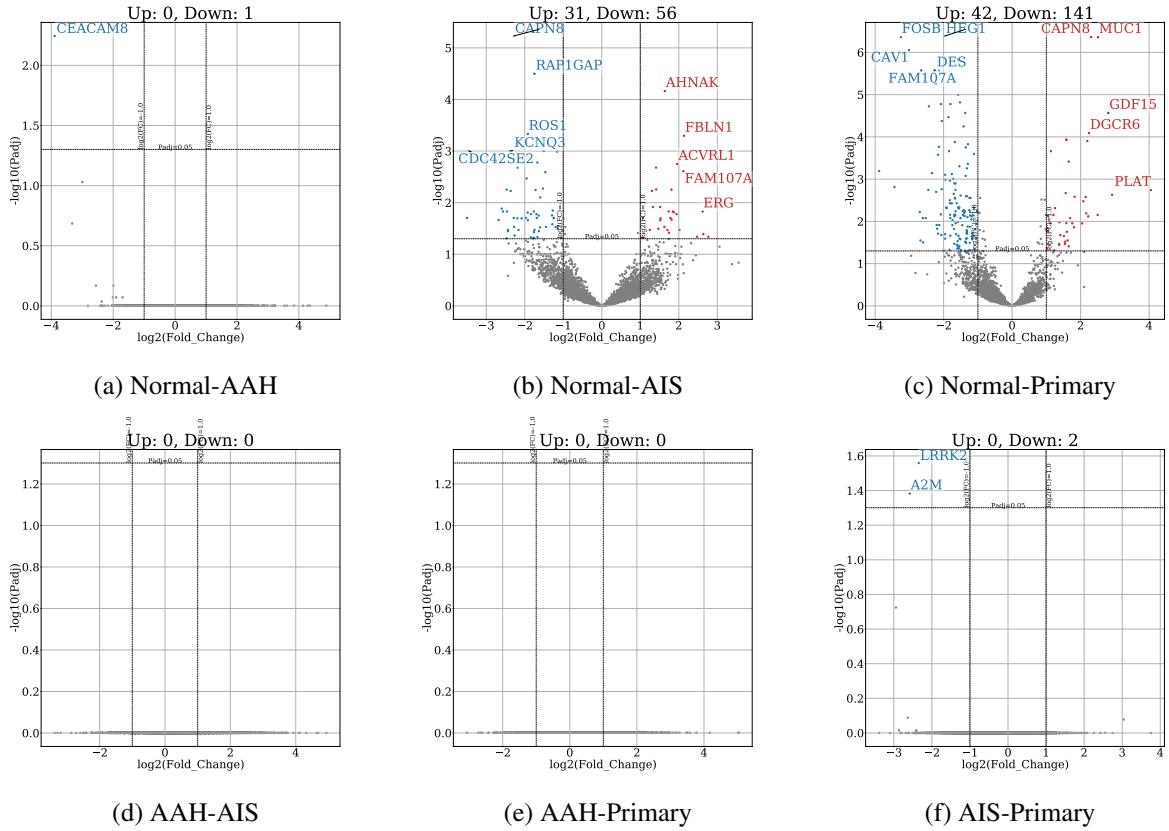


Figure 29: DEG volcano plots by STAR in ADC

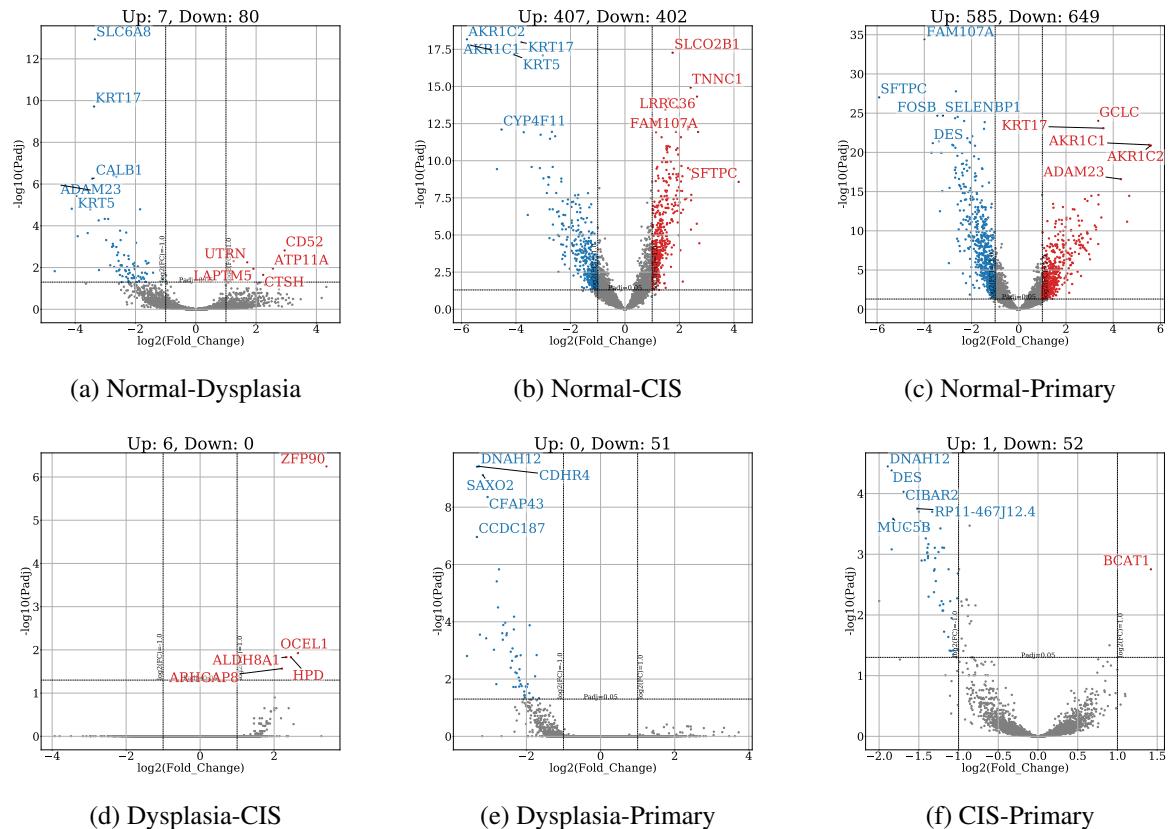


Figure 30: DEG volcano plots by Bowtie2 in SQC

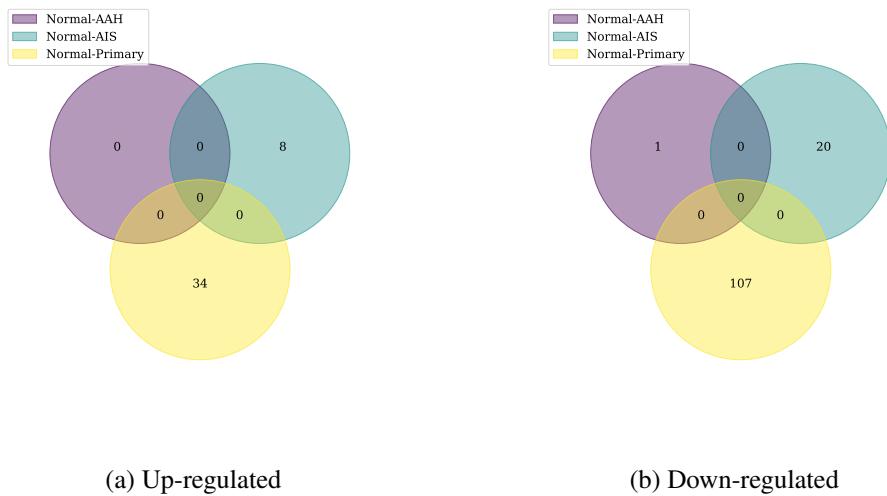


Figure 31: DEG Venn Diagram by Bowtie2 in ADC

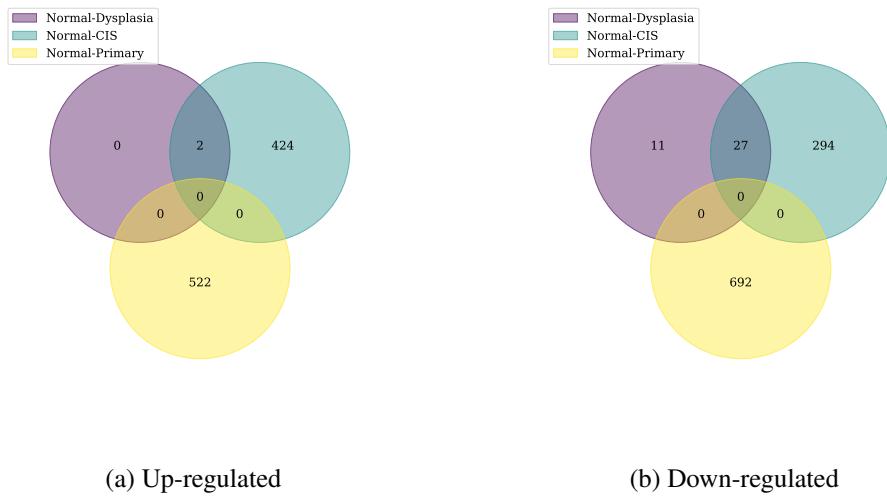


Figure 32: DEG Venn Diagram by Bowtie2 in SQC

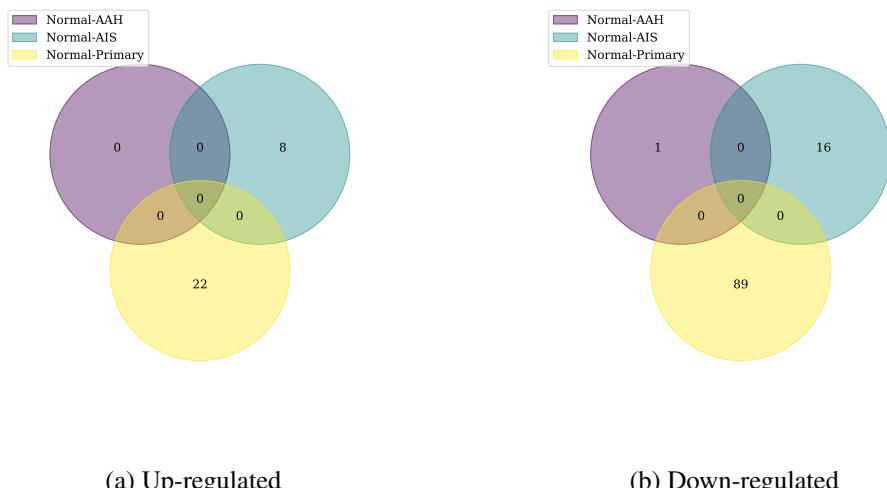


Figure 33: DEG Venn Diagram by STAR in ADC

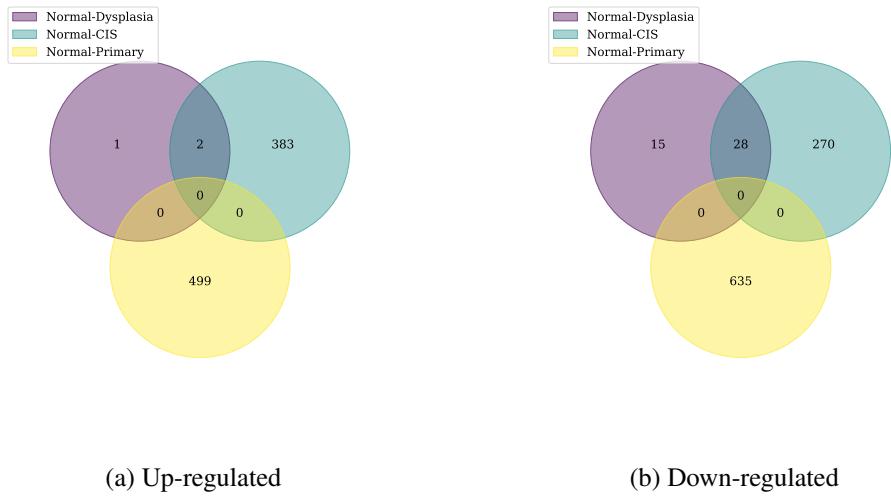


Figure 34: DEG Venn Diagram by STAR in SQC

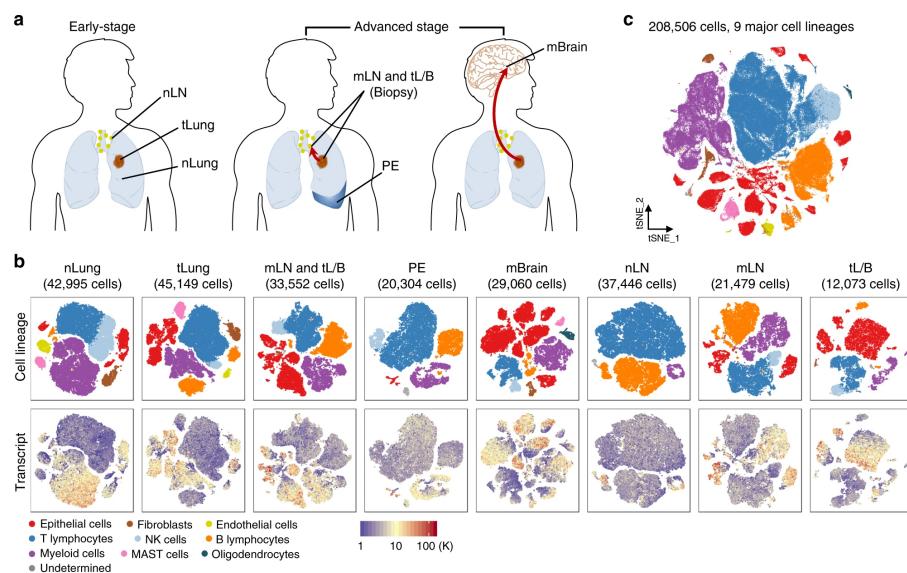


Figure 35: Comprehensive dissection and clustering of 208,506 single cells from LUAD patients (Kim et al., 2020)

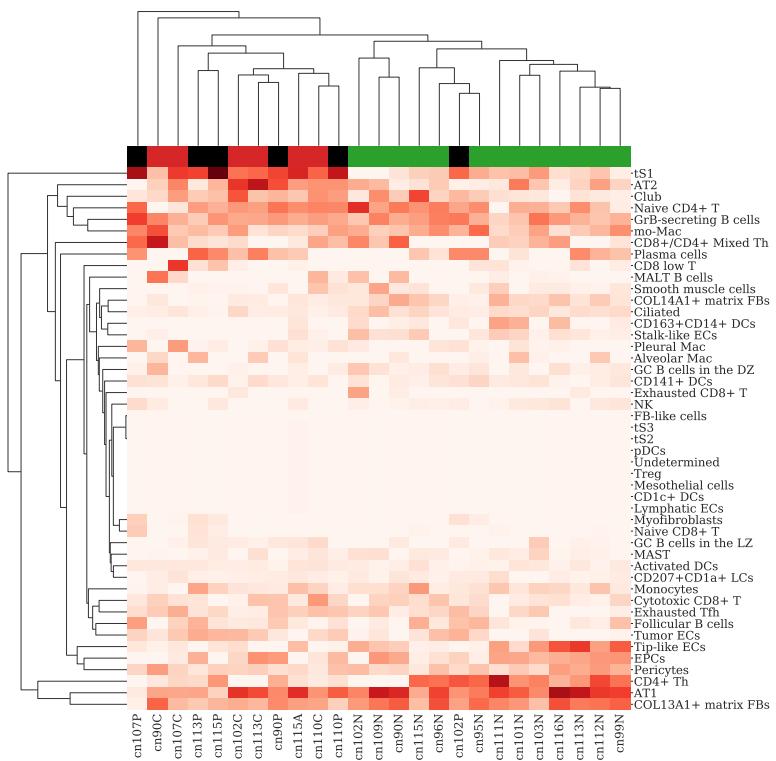


Figure 36: Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in ADC

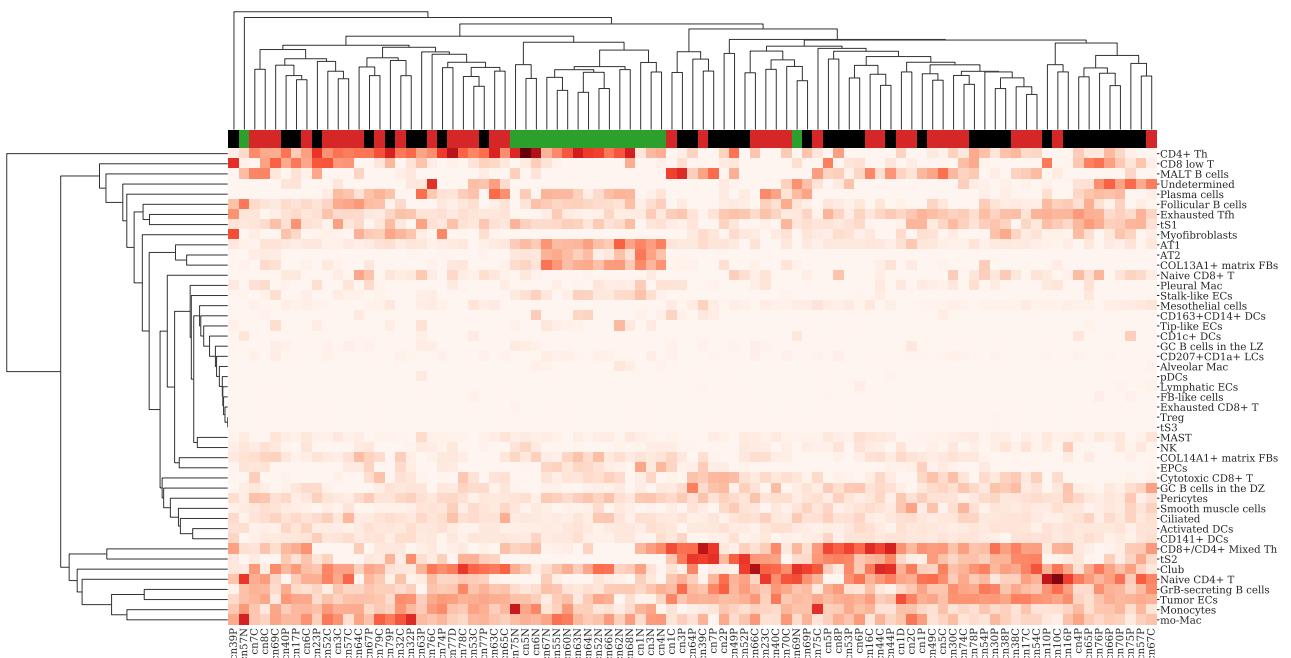


Figure 37: Cell deconvolution clustermap by Bowtie2 and CIBERSORTx in SQC

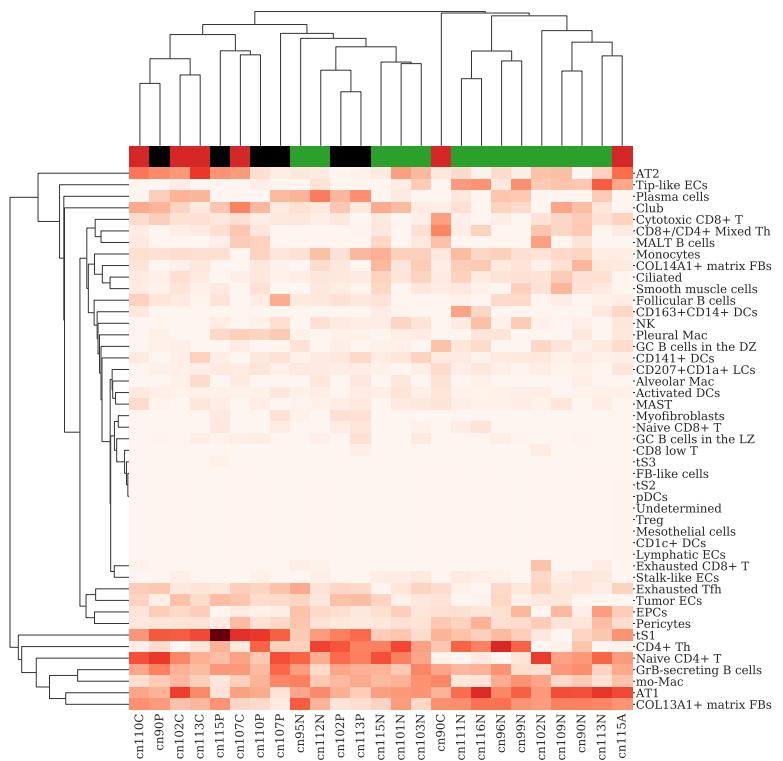


Figure 38: Cell deconvolution clustermap by STAR and CIBERSORTx in ADC

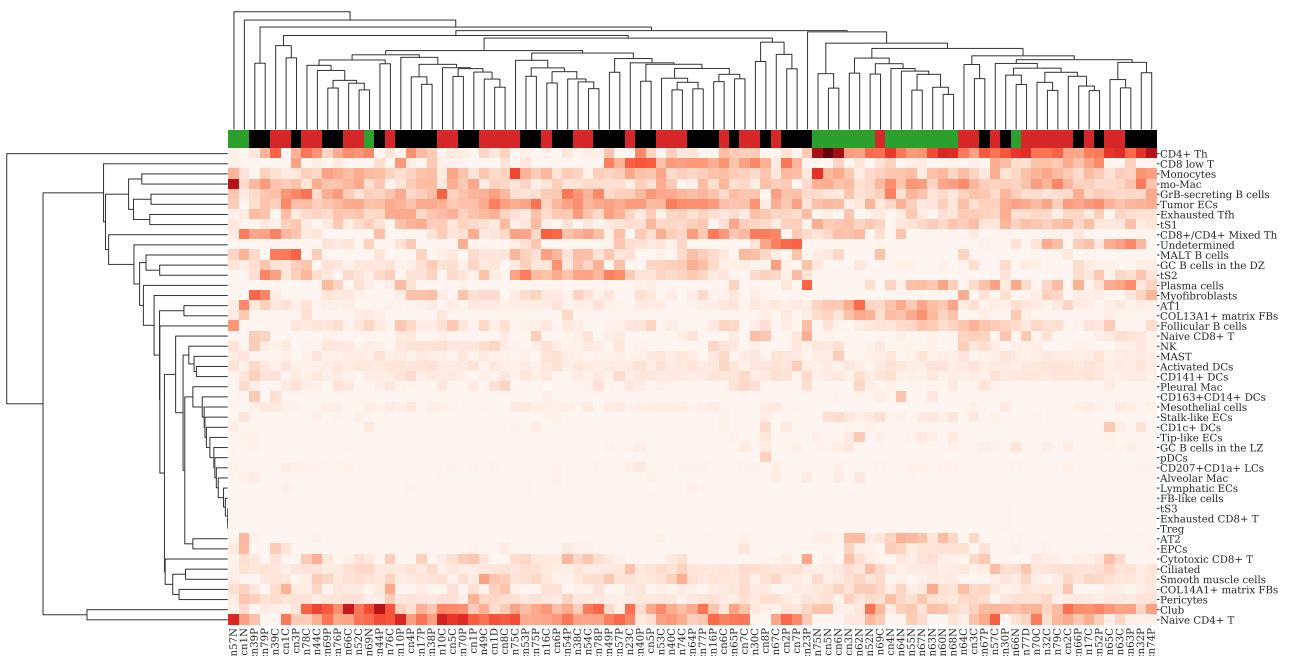


Figure 39: Cell deconvolution clustermap by STAR and CIBERSORTx in SQC

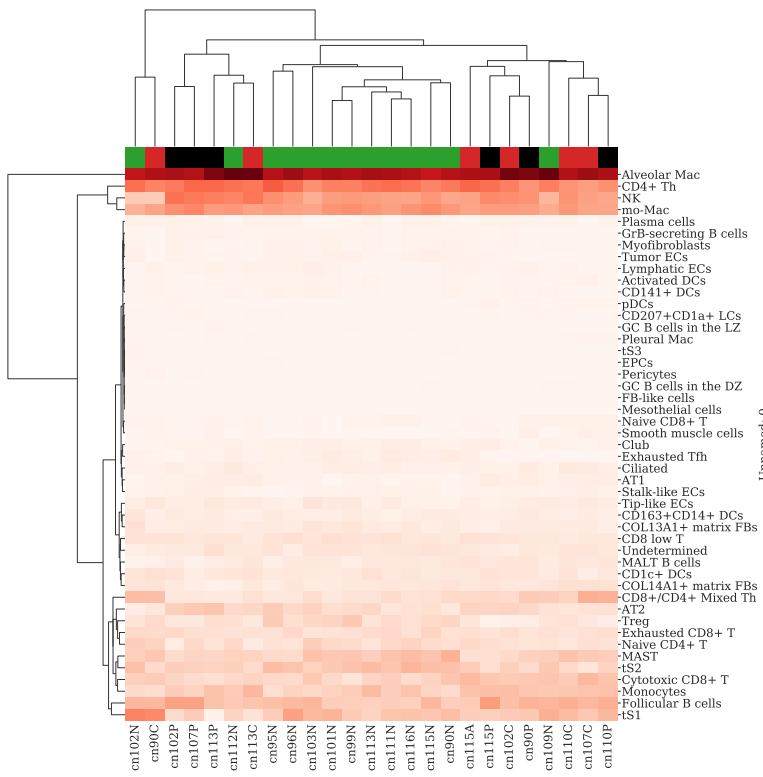


Figure 40: Cell deconvolution clustermap by Bowtie2 and BisqueRNA in ADC

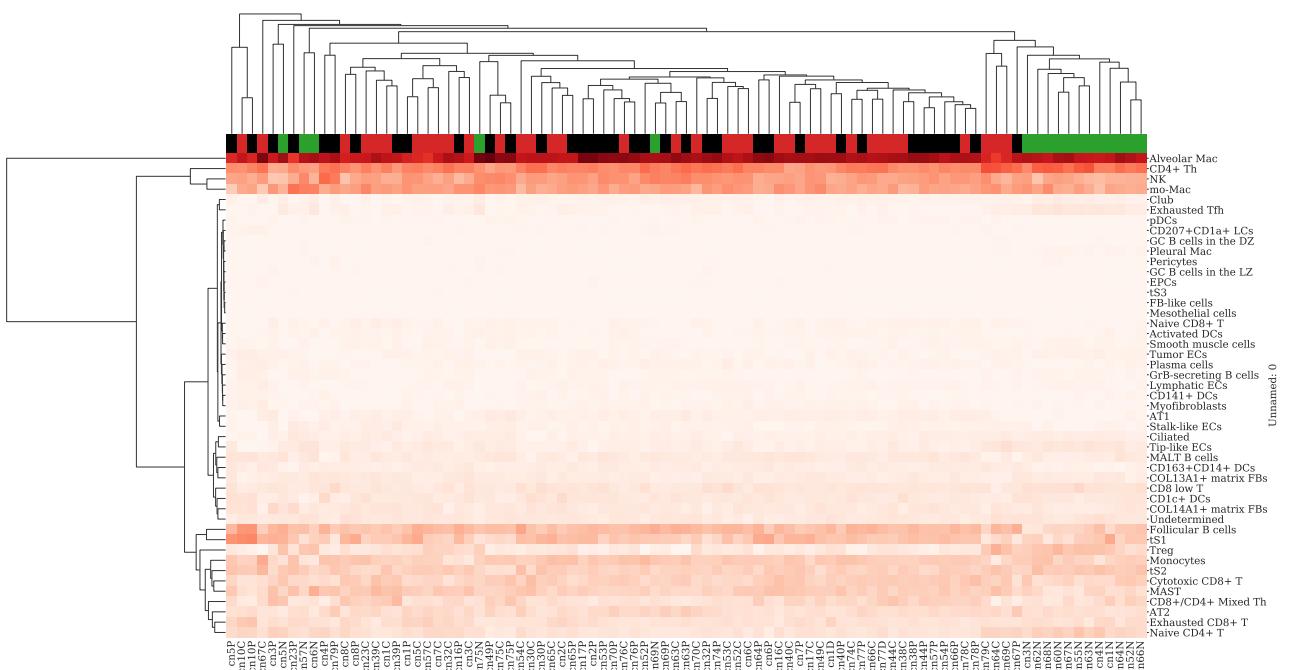


Figure 41: Cell deconvolution clustermap by Bowtie2 and BisqueRNA in SQC

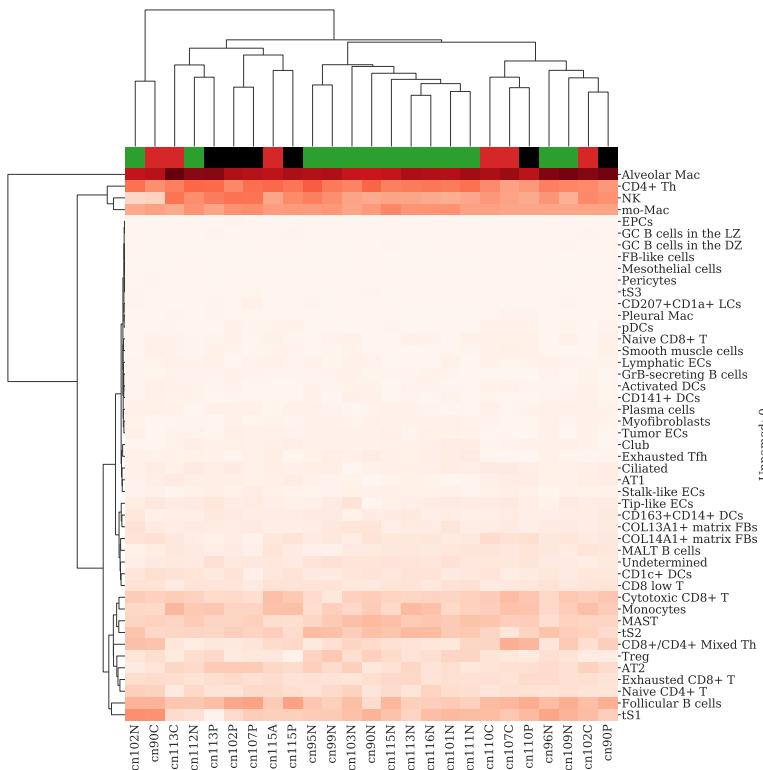


Figure 42: Cell deconvolution clustermap by STAR and BisqueRNA in ADC

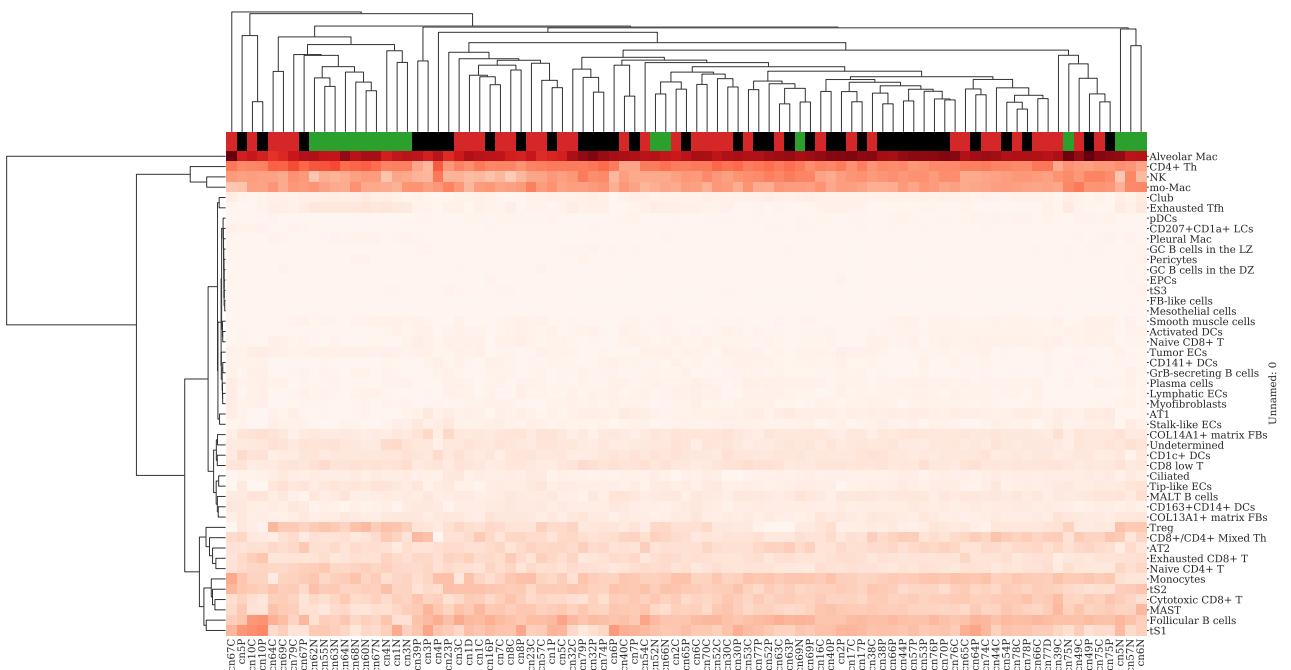


Figure 43: Cell deconvolution clustermap by STAR and BisqueRNA in SQC

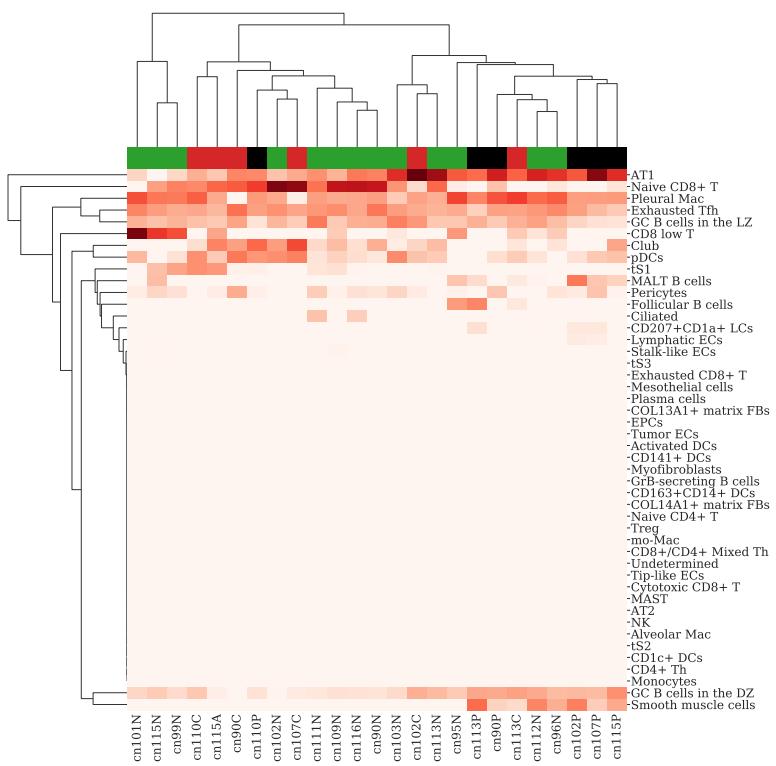


Figure 44: Cell deconvolution clustermap by Bowtie2 and MuSiC in ADC

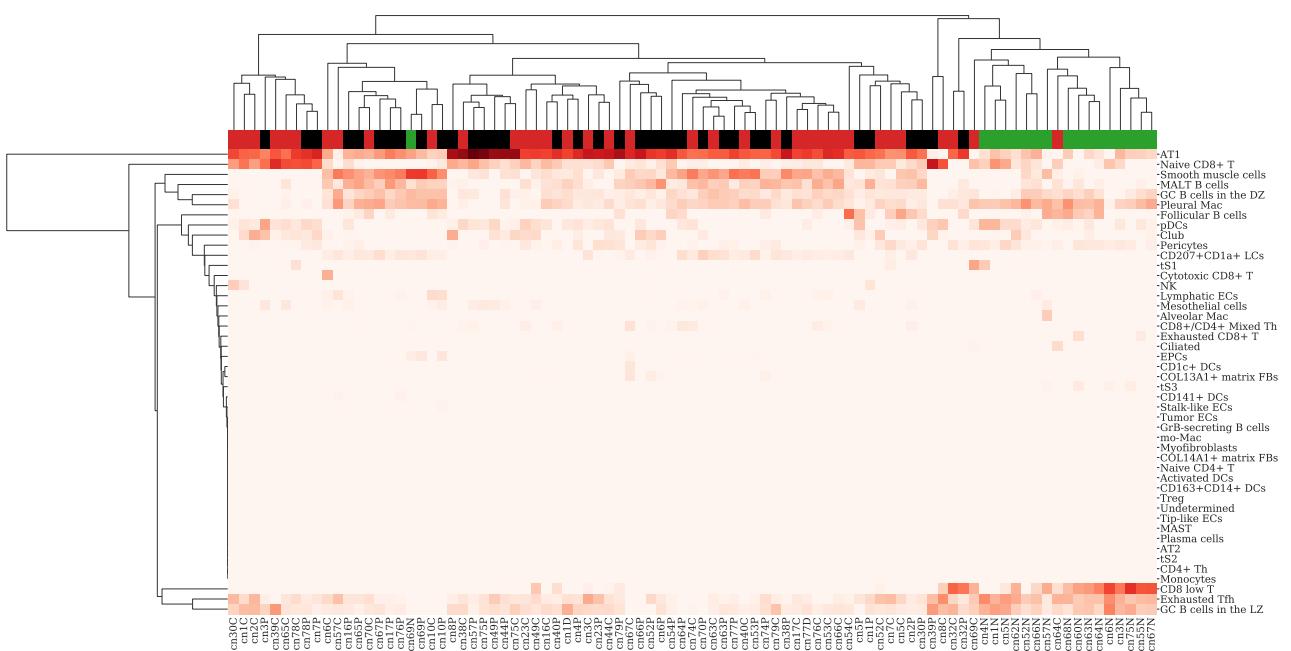


Figure 45: Cell deconvolution clustermap by Bowtie2 and MuSiC in SQC

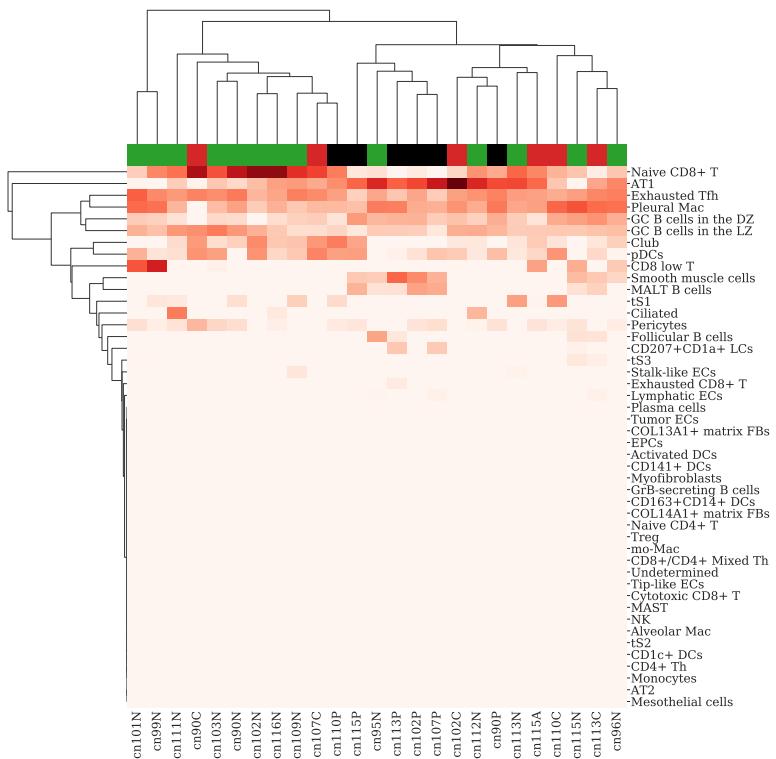


Figure 46: Cell deconvolution clustermap by STAR and MuSiC in ADC

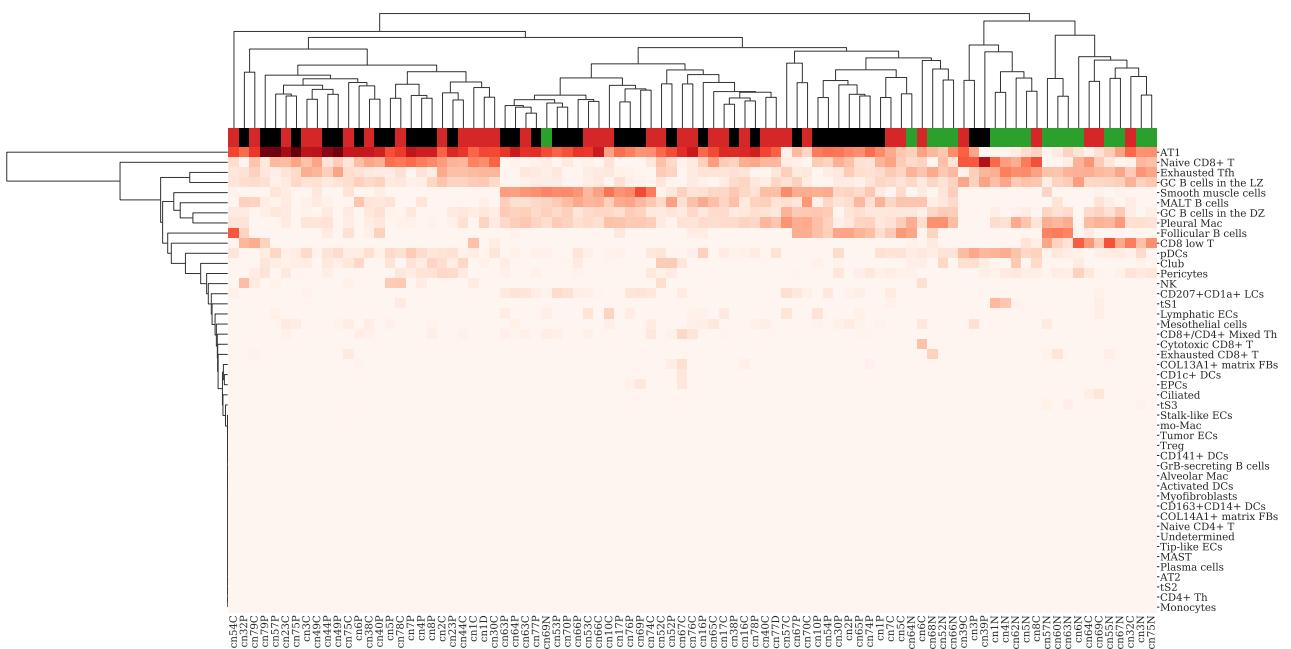


Figure 47: Cell deconvolution clustermap by STAR and MuSiC in SQC

#### 4.3.4 Findings in Copy Number Variation Analysis

### 4.4 Somatic Short Variation

#### 4.4.1 Somatic Short Variation Analysis with Mutect2

#### 4.4.2 Findings in Somatic Short Variation Analysis

### 4.5 Variant Allele Frequencies

### 4.6 Differences in Gene Expression levels

### 4.7 Bulk Cell Deconvolution

#### 4.7.1 Single-cell Reference Data

#### 4.7.2 CIBERSORTx

#### 4.7.3 BisqueRNA

#### 4.7.4 MuSiC

#### 4.7.5 SCDC

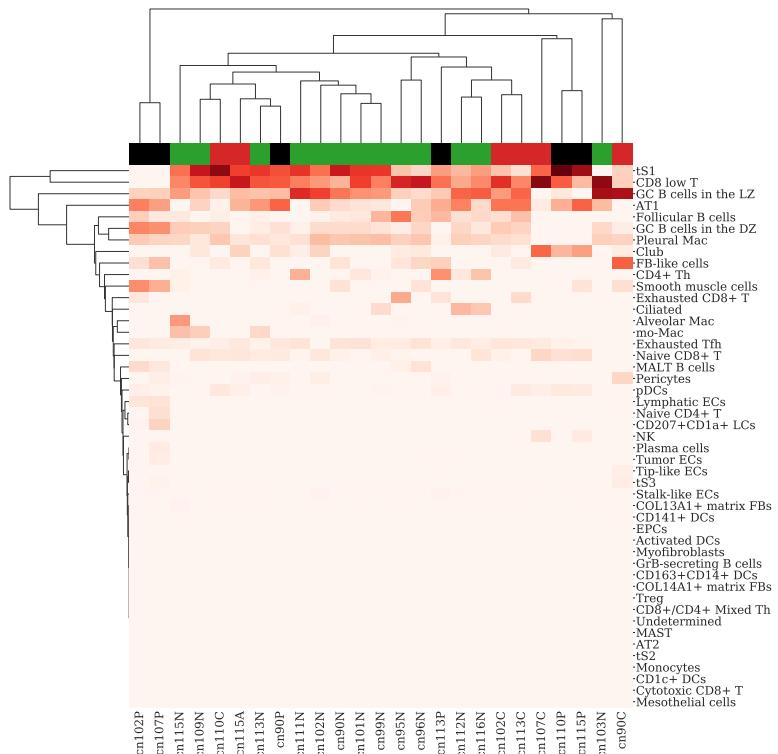


Figure 48: Cell deconvolution clustermap by Bowtie2 and SCDC in ADC

## 5 Discussion

## 6 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.

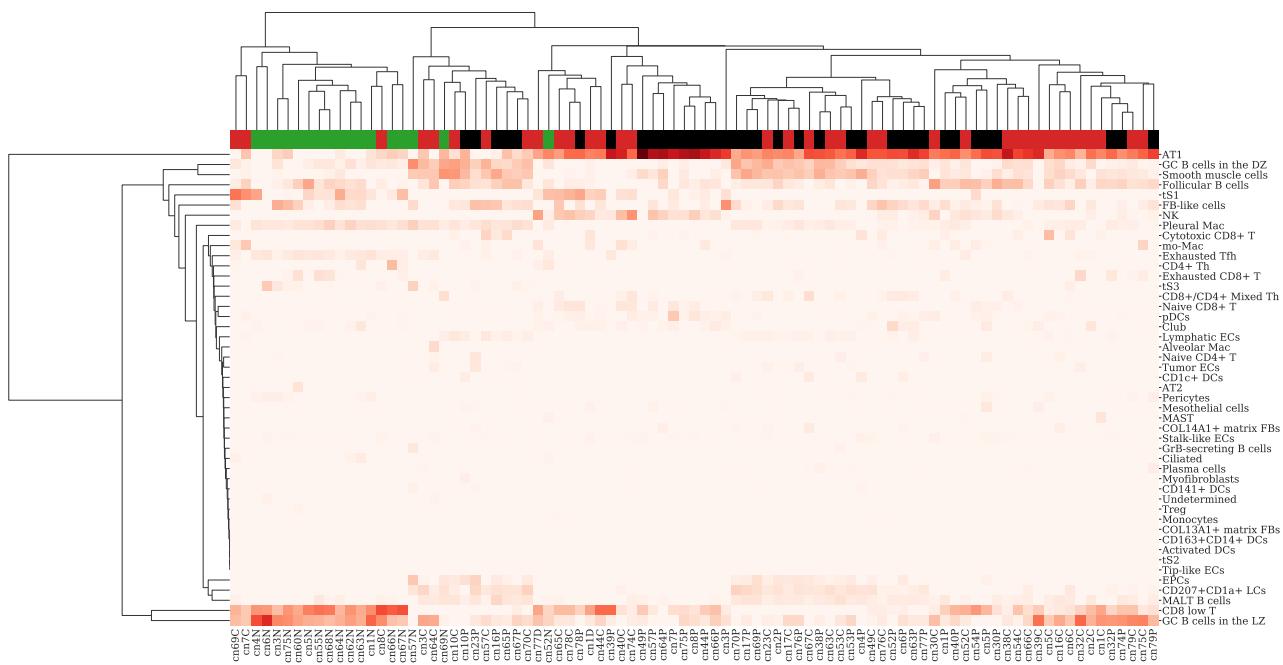


Figure 49: Cell deconvolution clustermap by Bowtie2 and SCDC in SQC

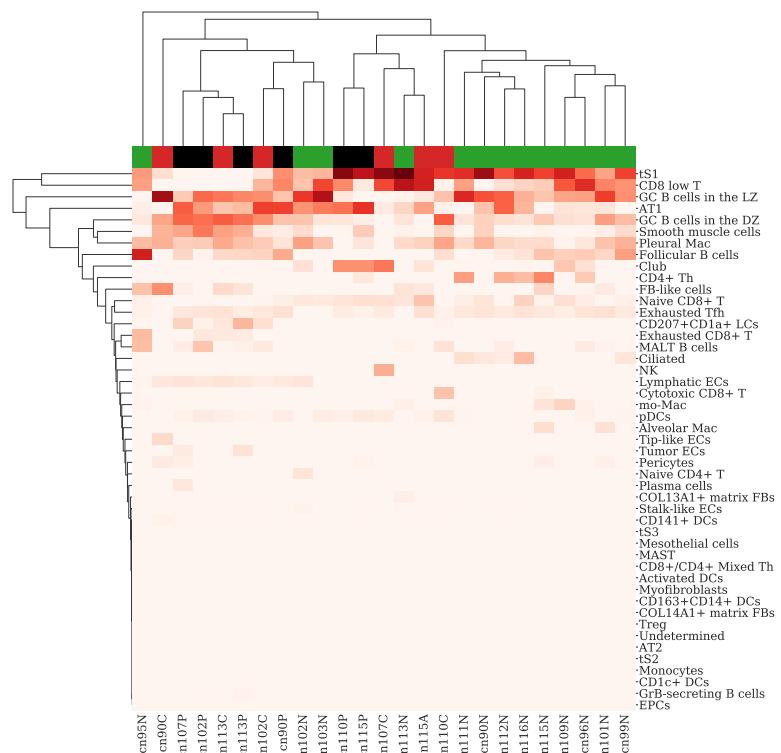


Figure 50: Cell deconvolution clustermap by STAR and SCDC in ADC

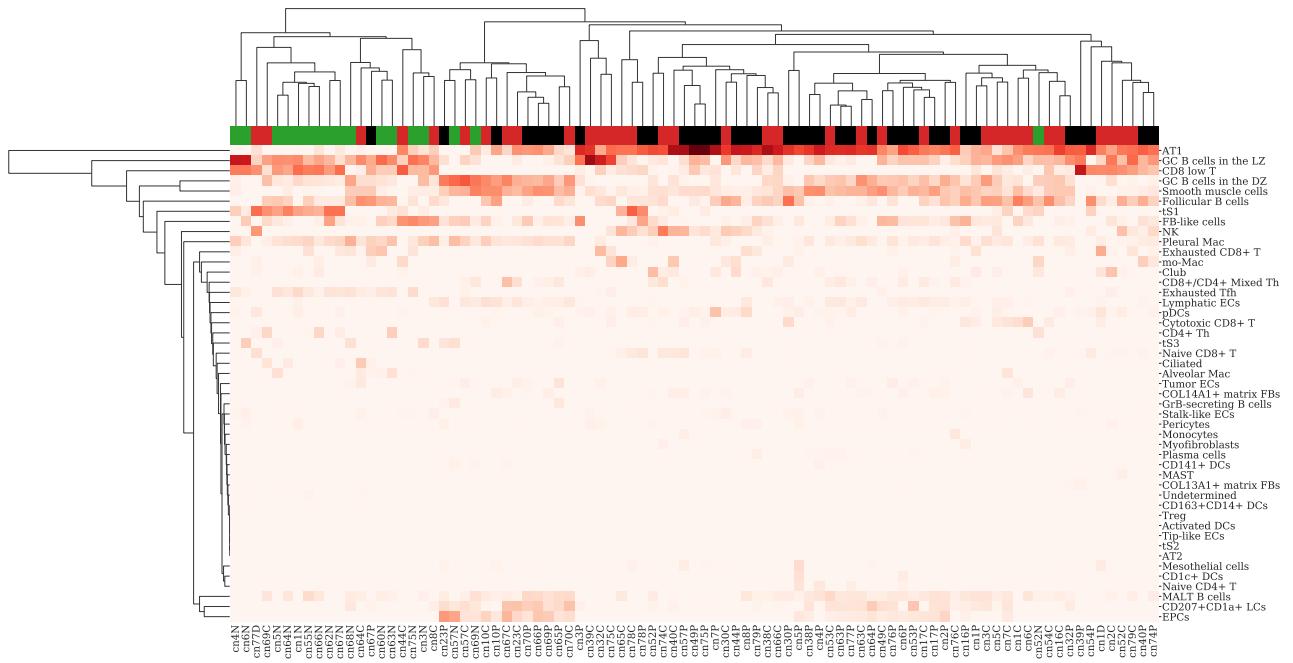


Figure 51: Cell deconvolution clustermap by STAR and SCDC in SQC

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.

Kim, N., Kim, H. K., Lee, K., Hong, Y., Cho, J. H., Choi, J. W., ... others (2020). Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications*, 11(1), 1–15.

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.