Lung Precancer Study

Jaewoong Lee S. Park Y. Choi I. Yun Semin Lee

Department of Biomedical Engineering
Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2021-09-16

Overview

1. Introduction

1. Introduction

1.1. Lung Cancer

Lung Cancer?

The most common cancer

The most common form of cancer:
12.3 % of all cancers (Minna, Roth, & Gazdar, 2002)

The most important factor

Tobacco

Cancer Survival Rate in Korea

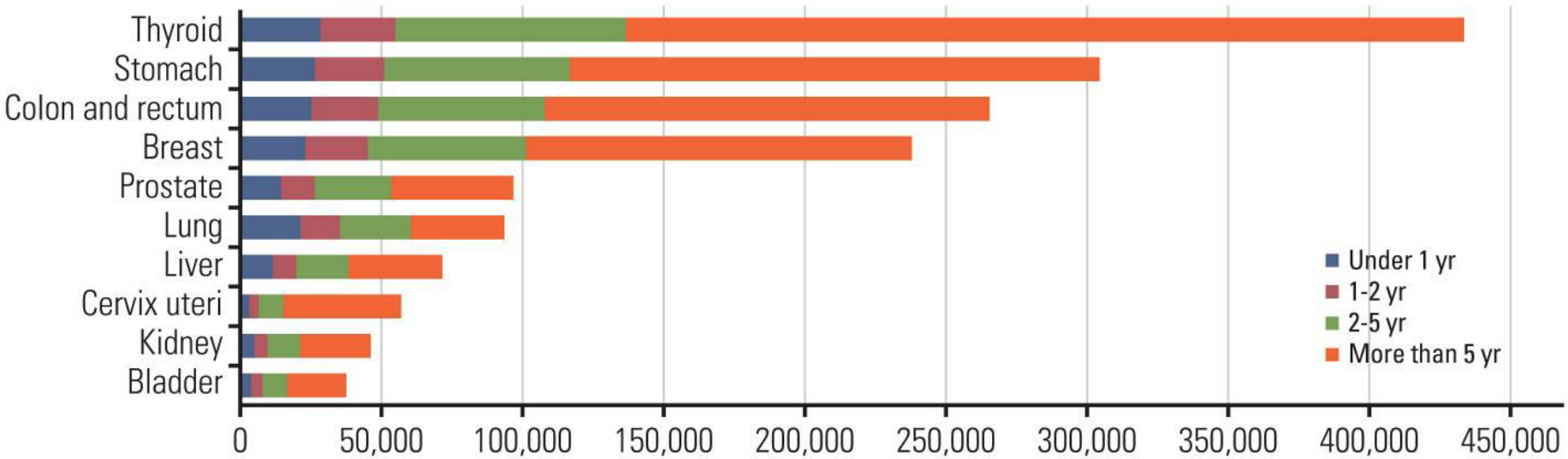


Figure: Common cancer survival rates (Hong et al., 2021)

Survival rate (More than 5 yr)

- Tyroid: 68.4 %
- Lung: 35.4 %

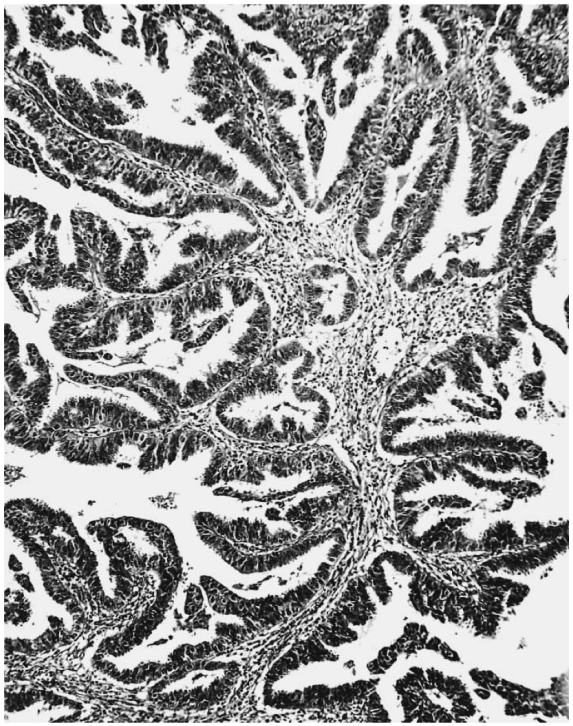
Type of Lung Cancer

Types of lung cancer:

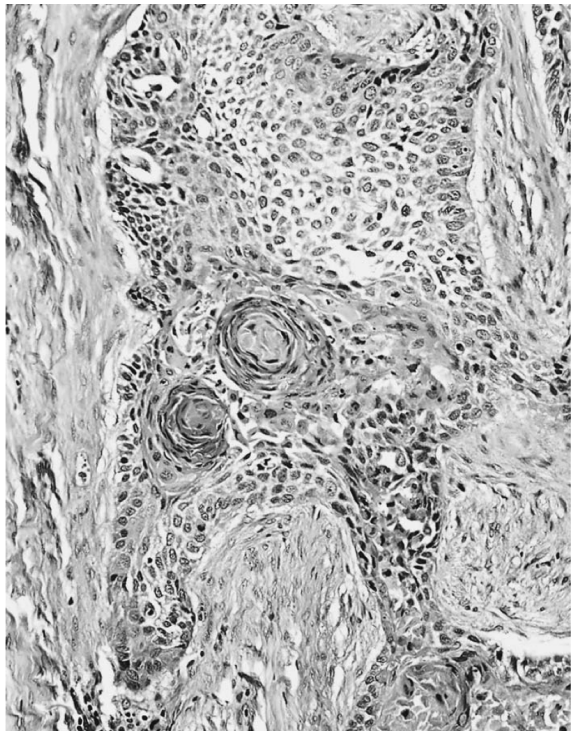
1. Adenocarcinoma (ADC) (40 %) ★
2. Squamous cell carcinoma (SQC) (25 %) ★
3. Small cell carcinoma (20 %)
4. Large cell carcinoma (10 %)
5. Adenosquamous carcinoma (< 5 %)
6. Carcinoid (< 5 %)
7. Bronchioalveolar (Bronchial gland carcinoma)

(Vincent et al., 1977; Collins, Haines, Perkel, & Enck, 2007)

ADC vs. SQC I



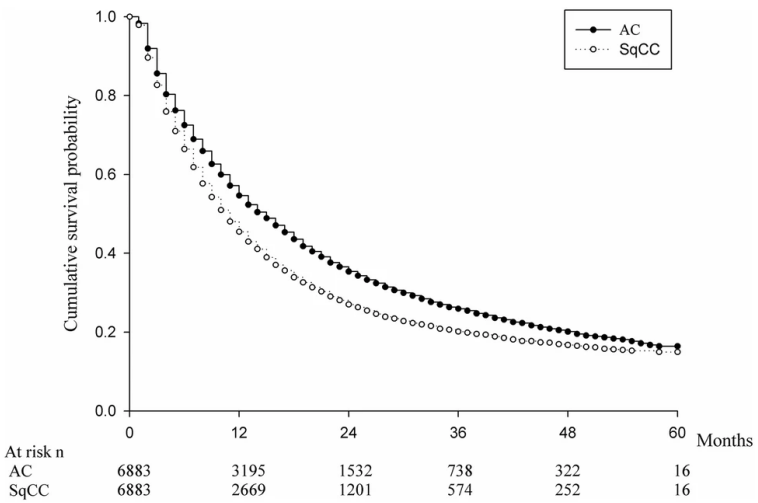
(a) ADC



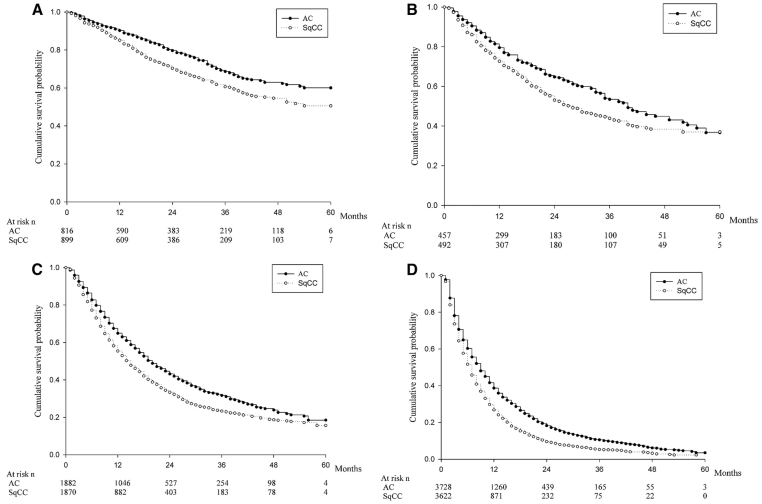
(b) SQC

Figure: ADC and SQC histology in Lung cancer (Travis, 2002)

ADC vs. SQC II



(a) All patients



(b) By cancer stages

Figure: Kaplan-Meiere survival curves for ADC & SQC (B.-Y. Wang et al., 2020)

Findings

SQC is more dangerous than ADC. $\because p < 0.001$

1. Introduction

1.2. Study Objectives

Study Objectives

Find different mutations

- between WES vs. WTS
- from cancer vs. precancer

Pathway examine

- with the mutation of WES & RNA-seq
- with immune-depleted animal models

Ultra-deep sequencing

to find an *infinitesimal* quantity of Non-Circulating Tumor DNA

- from blood
- from urine
- from bronchus

2. Materials

Lung Cancer Data

- WES (n=289) + Transcriptome (n=166)
- Normal + {Primary, CIS + AIS, AAH, Dysplasia, MIA}
  - Carcinoma in situ
  - Adenocarcinoma in situ
  - Atypical adenomatous hyperplasia
  - Dysplasia
  - Minimally invasive adenocarcinoma
- Squamouse cell carcinoma (SQC) & Adenocarcinoma (ADC)
  1. Normal $\rightarrow$ Dysplasia $\rightarrow$ CIS $\rightarrow$ SQC (n=80)
  2. Normal $\rightarrow$ AAH $\rightarrow$ AIS $\rightarrow$ MIA $\rightarrow$ ADC (n=28)

3. Methods

3. Methods

3.1. Workflows

Data pre-processing for variant discovery

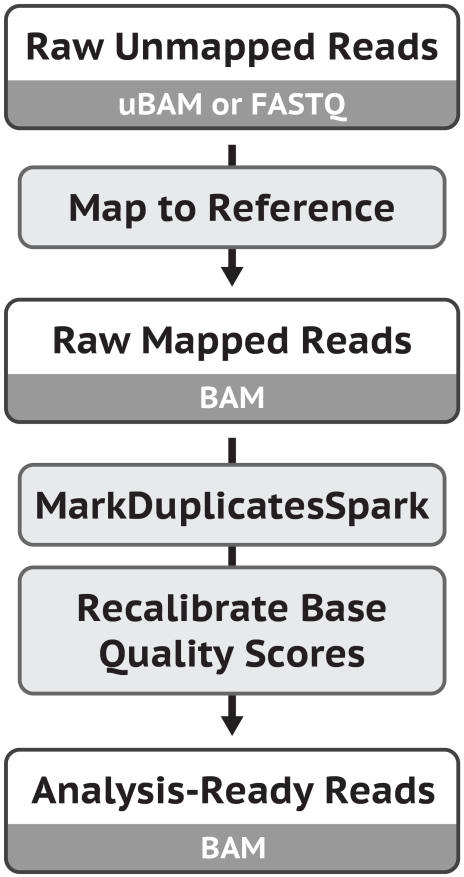


Figure: Data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

Somatic short variant discovery

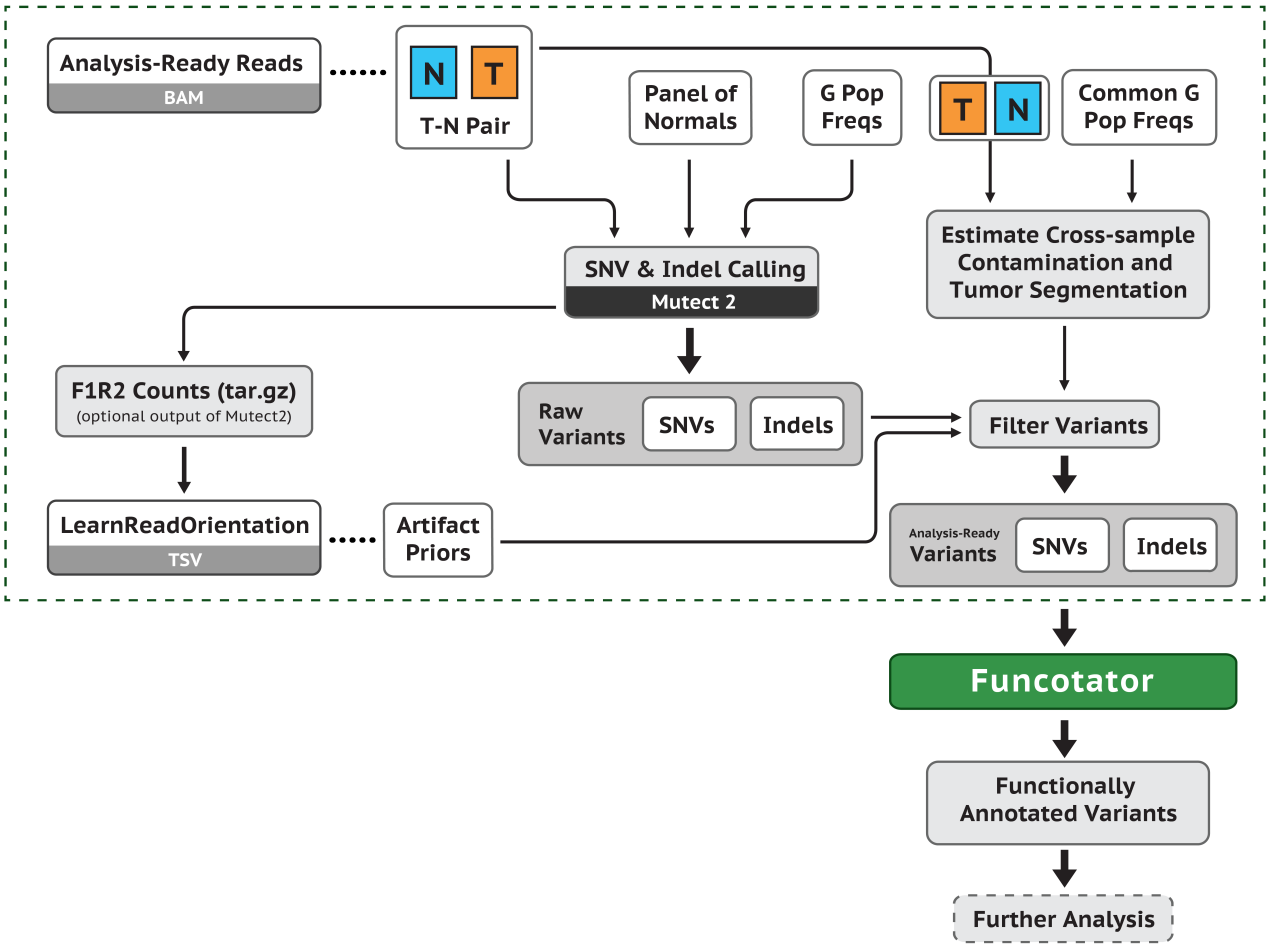


Figure: Somatic short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

Germline short variant discovery

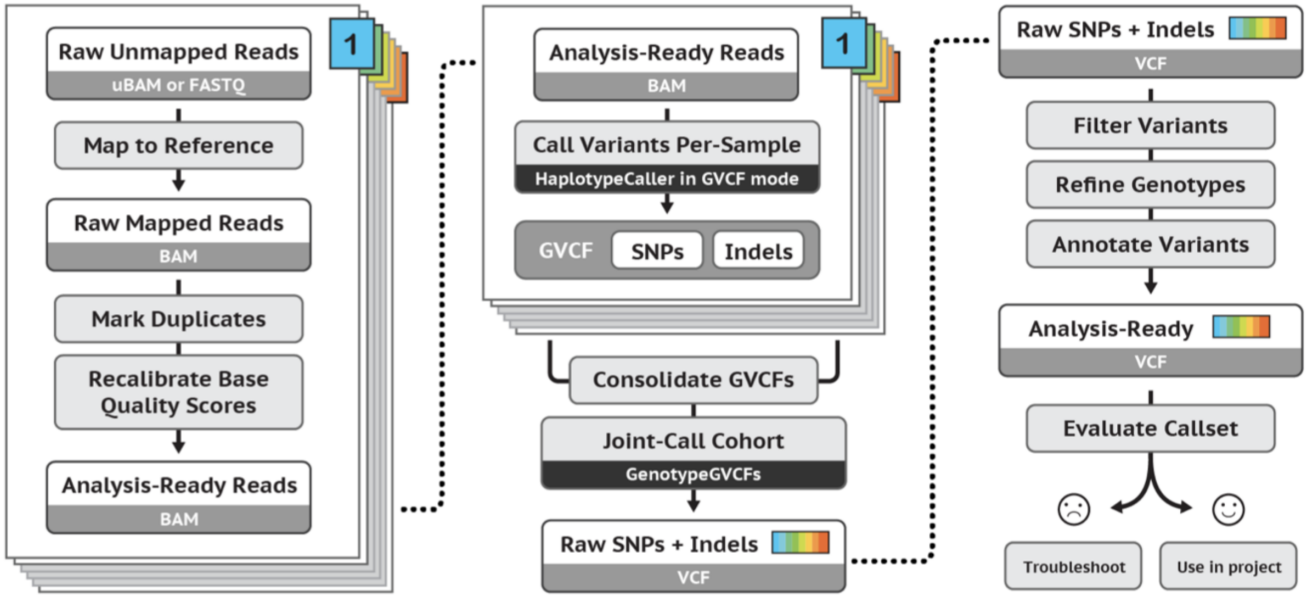


Figure: Germline short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

RNA-seq short variant discovery

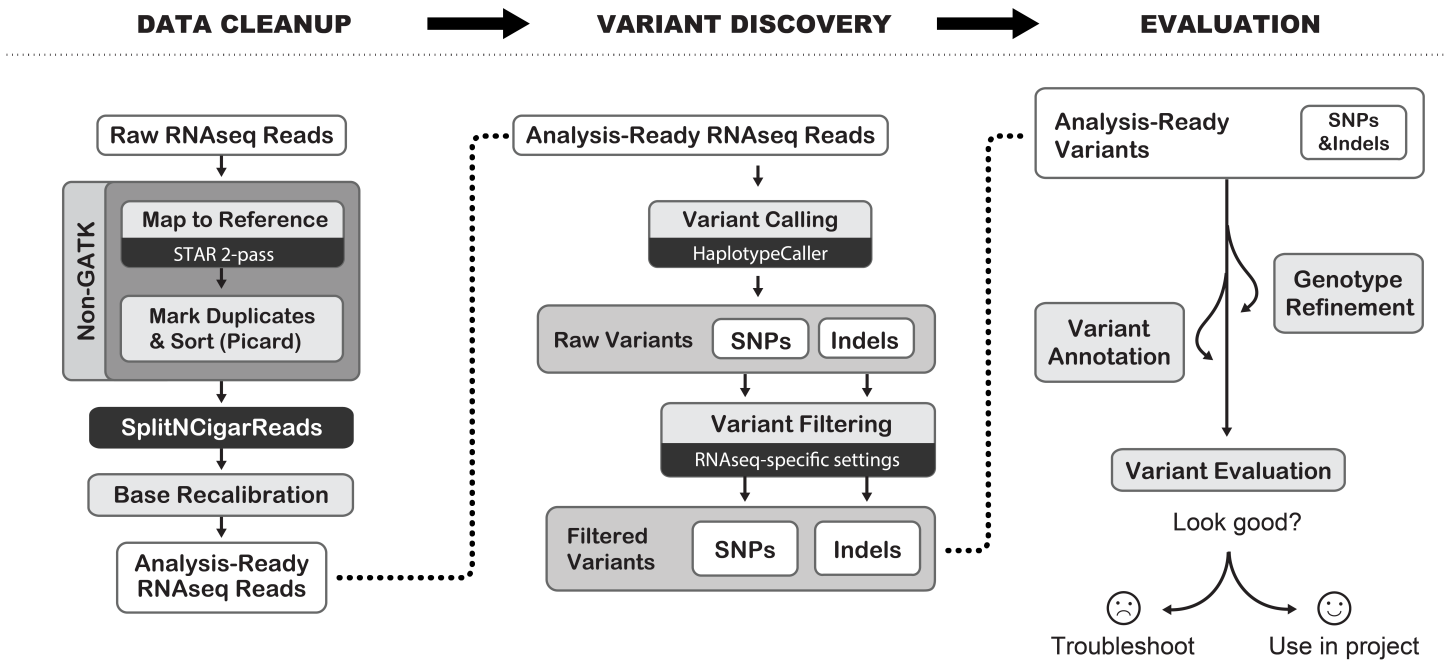


Figure: RNA-seq short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

4. Results

4. Results

4.1. Quality Checks

FastQC?

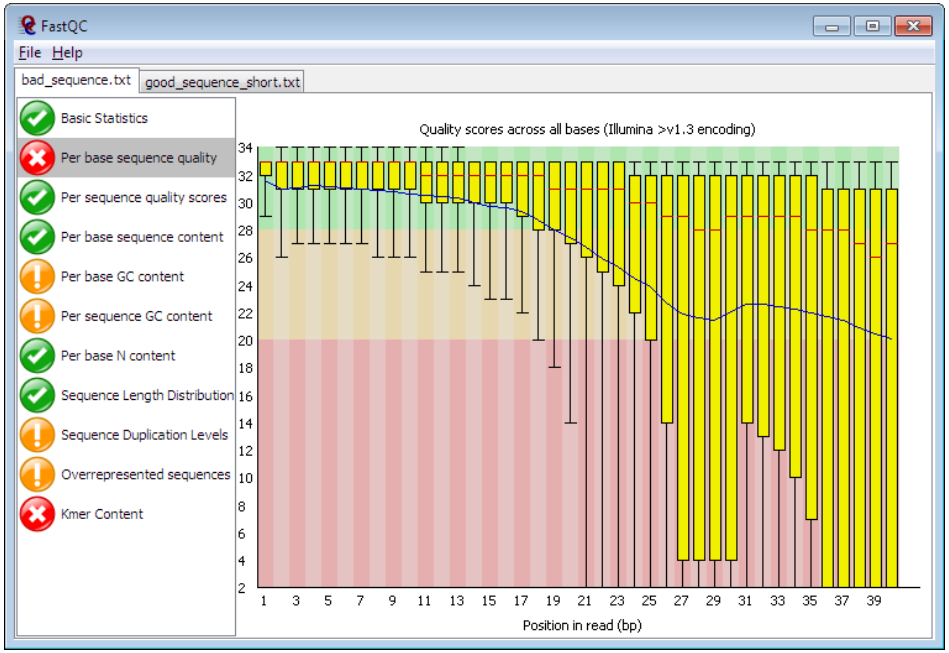


Figure: Example of FastQC Result (Andrews et al., 2012)

- A quality check tool for sequence data
- Give an overview that which test may be problems

FastQC on WES

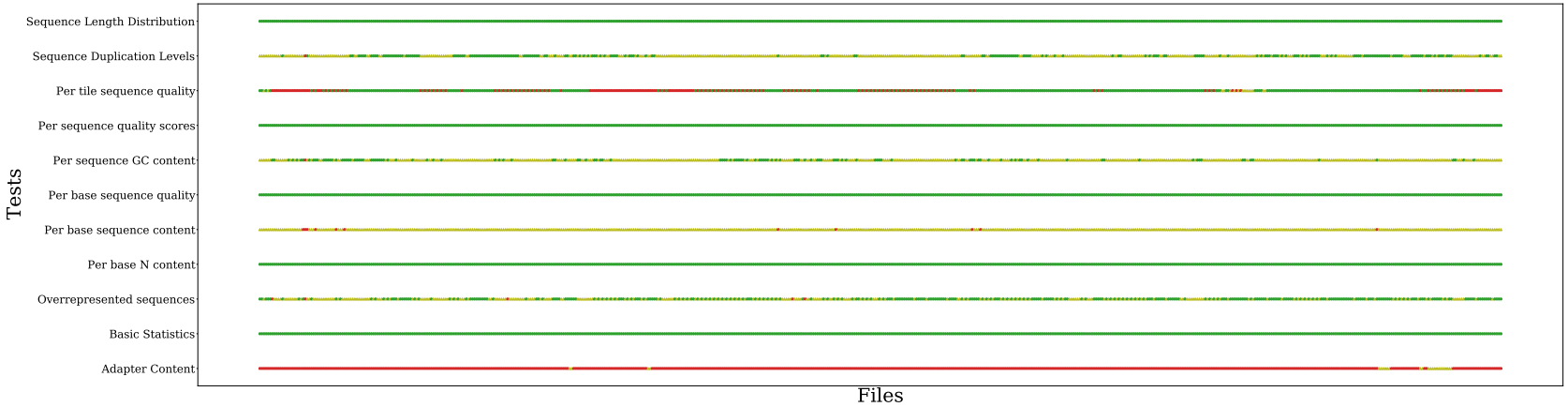
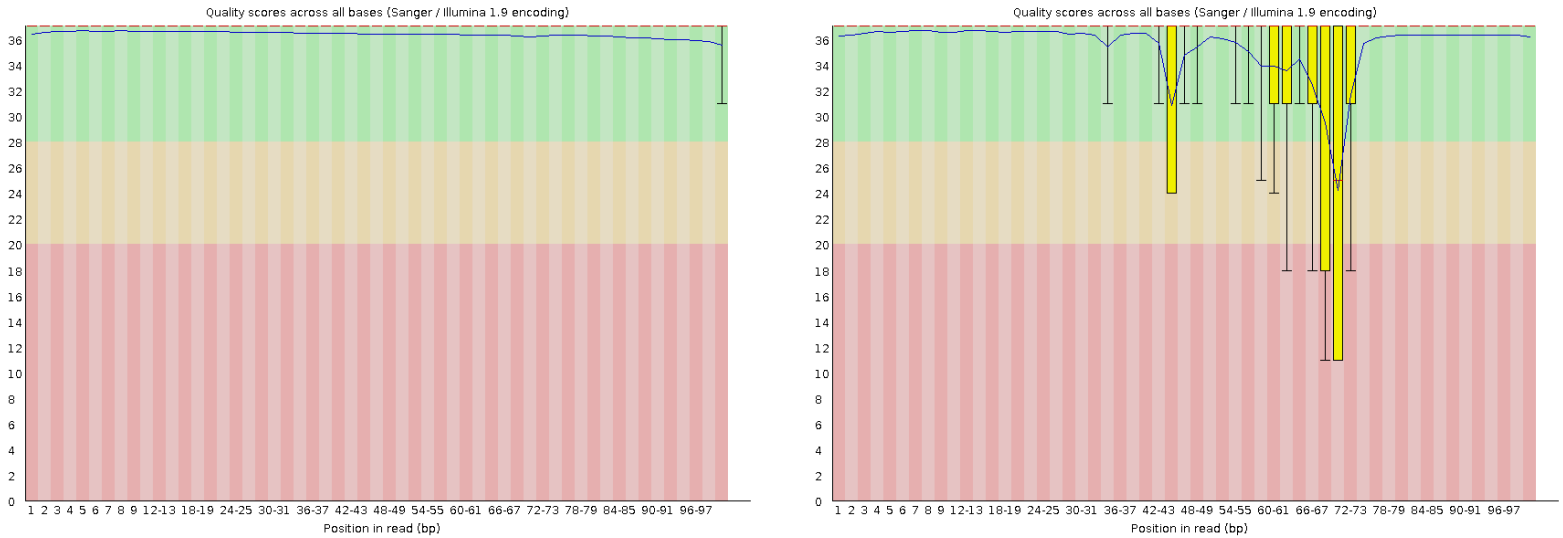


Figure: FastQC with WES data

Failure on 33P1 sample

33P1 is excluded at further analysis.

Failure on 33P1 I



(a) 33N (b) 33P1

Figure: Per Base Sequence Quality Results

Failure on 33P1 II

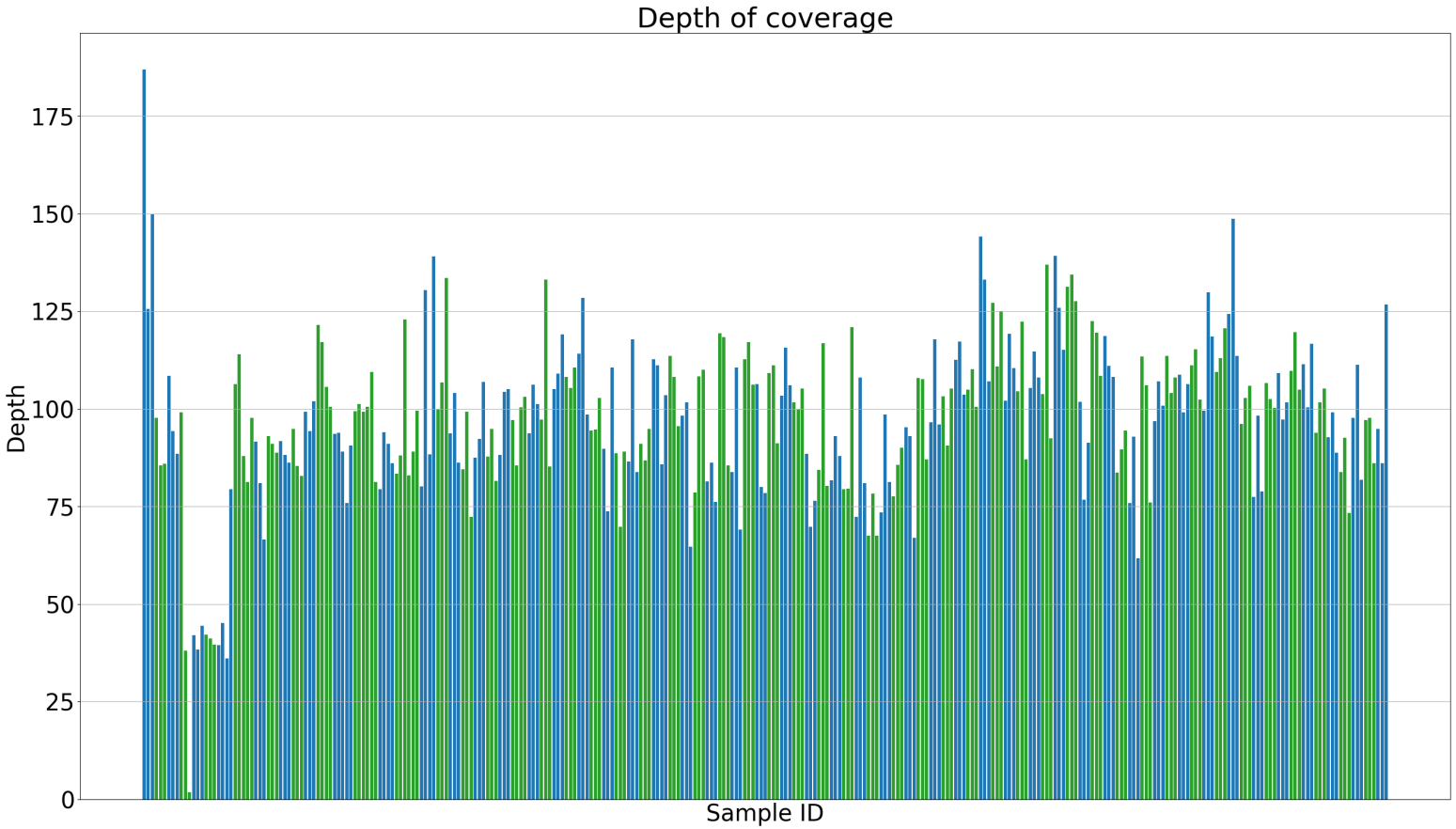


Figure: Coverage Depth Plot

FastQC on WTS

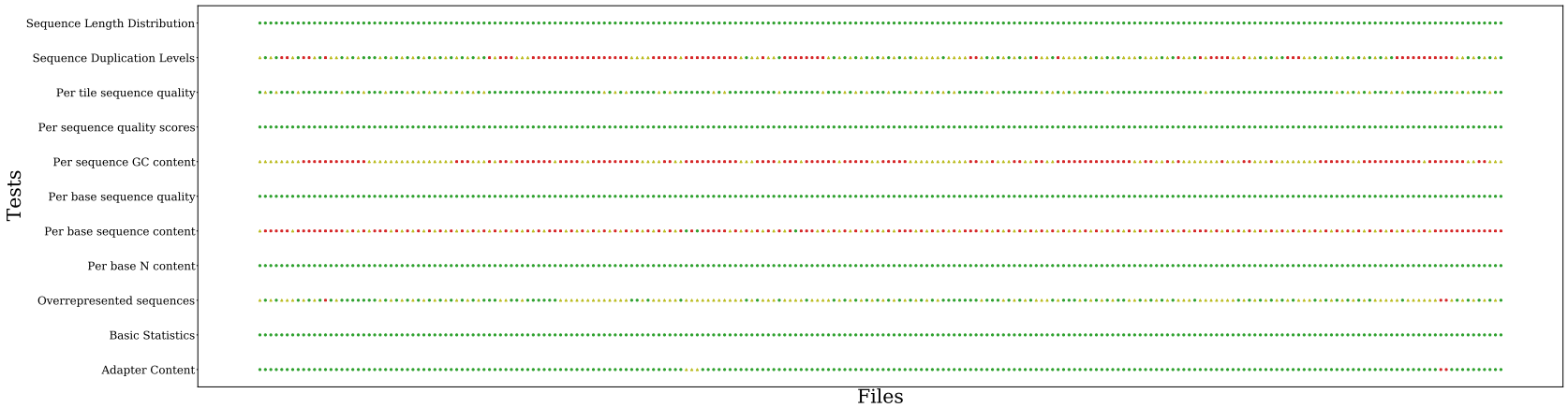


Figure: FastQC with WTS data

All sample are good to analysis

∵ No sample has more than 5 failures.

4. Results

4.2. Copy Number Variations

Sequenza?

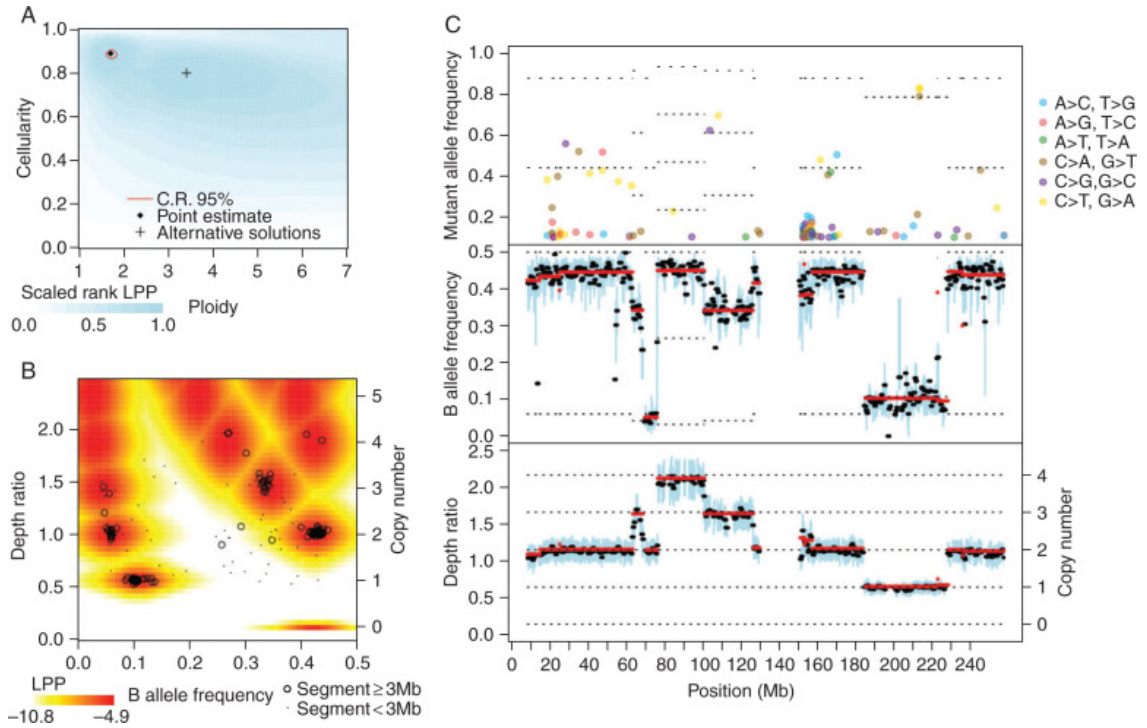
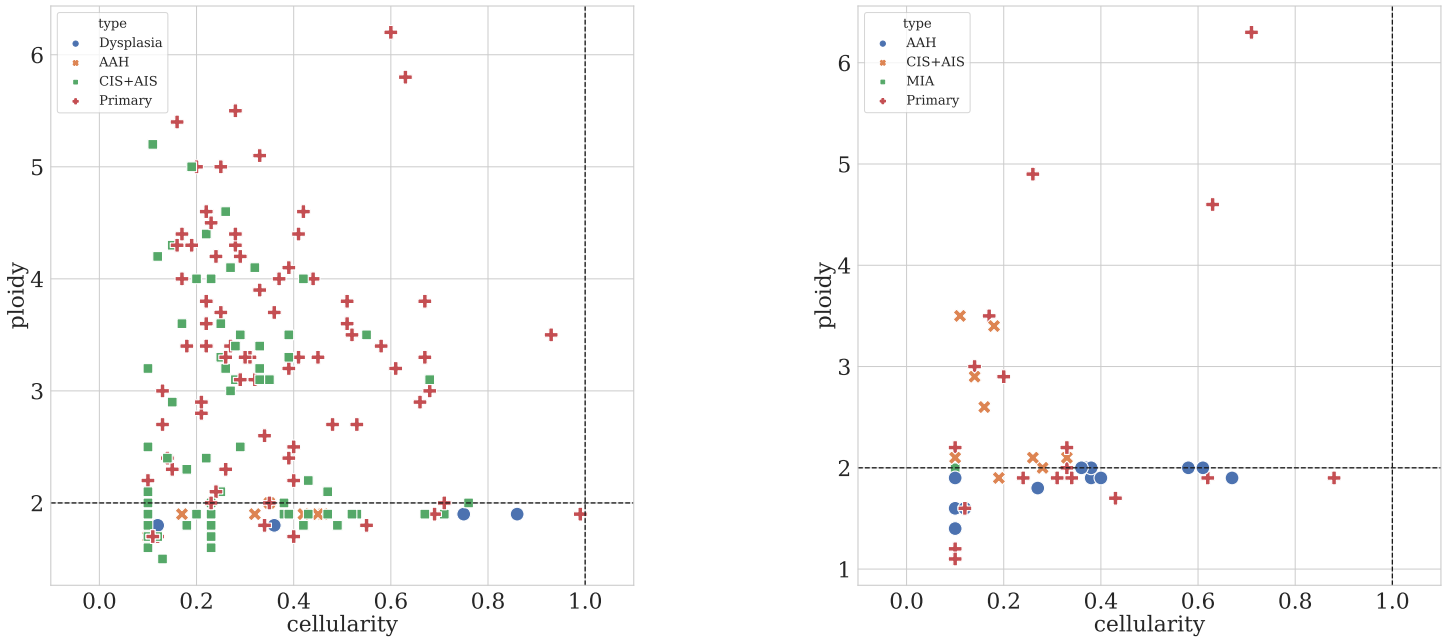


Figure: Representative Output of the Sequenza (Favero et al., 2015)

Cellularity & Ploidy on WES



(a) SQC Samples

(b) ADC Samples

Figure: Cellularity and Ploidy from Sequenza

Genome View on Patient #57

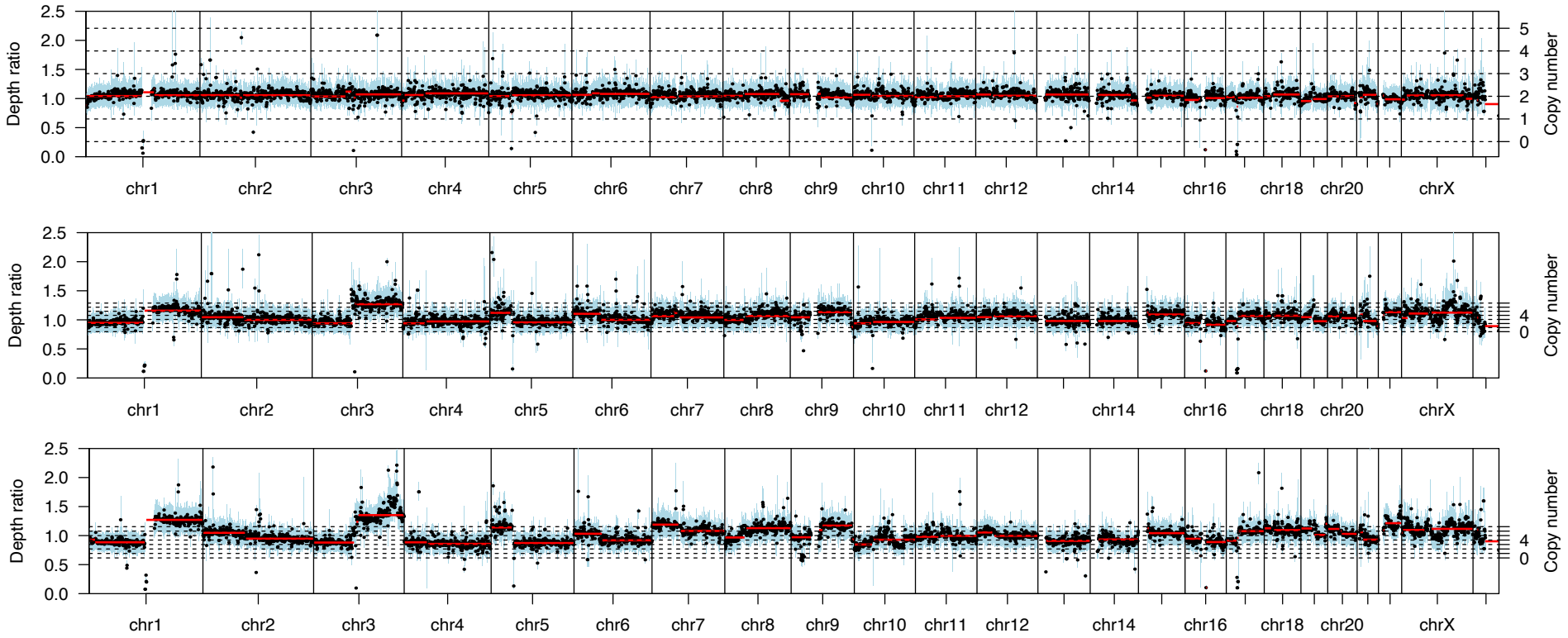


Figure: Dysplasia-CIS-Primary Tumor on Patient #57

CNVs of SQC

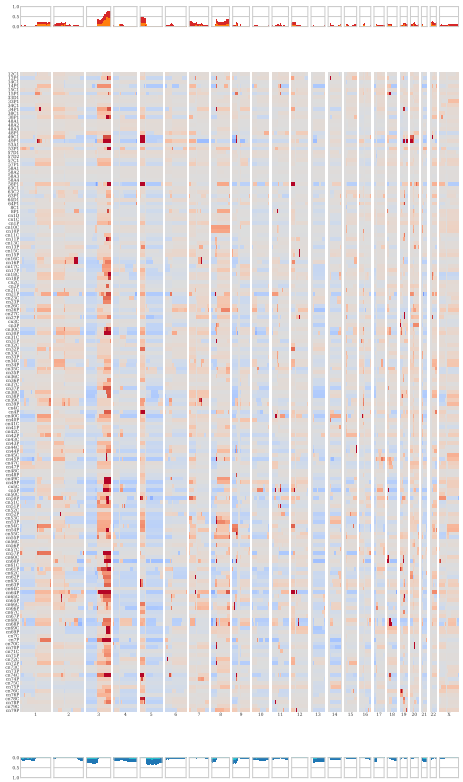


Figure: CNV Plot with SQC Patients

CNVs of ADC

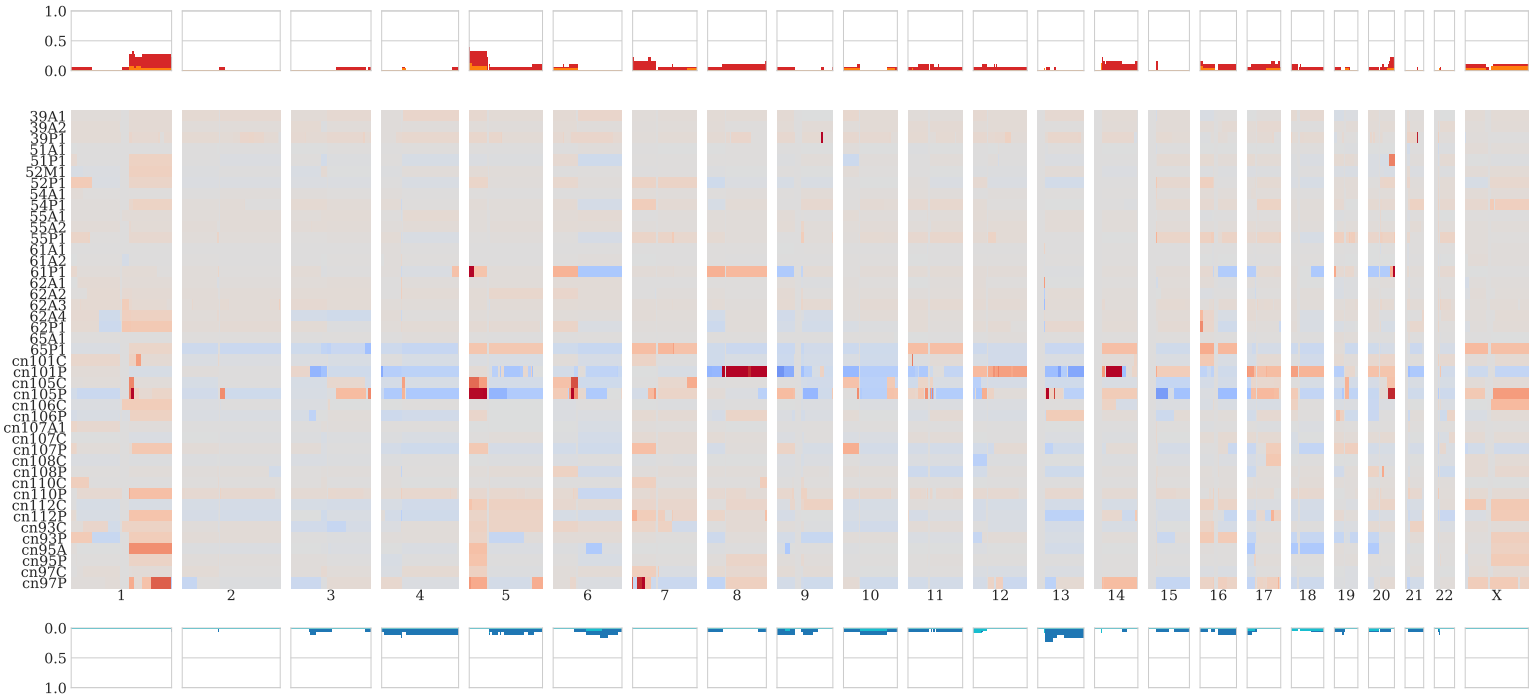


Figure: CNV Plot with ADC Patients

SQC vs. ADC in CNV Plot

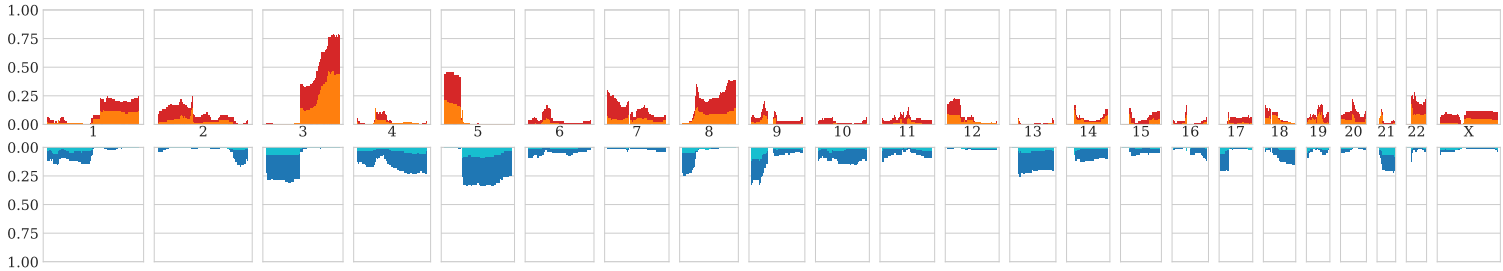


Figure: Simple CNV Plot with SQC Patients

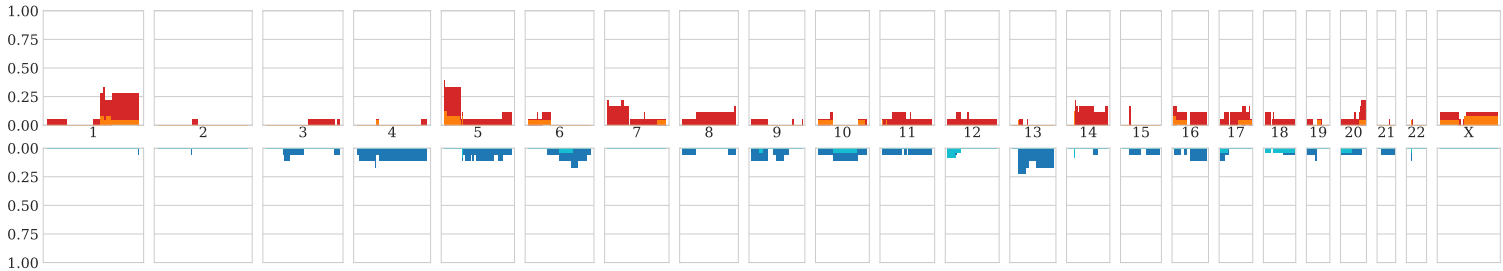


Figure: Simple CNV Plot with ADC Patients

Findings in Sequenza

4. Results

4.3. SNVs Analysis

Mutect2?

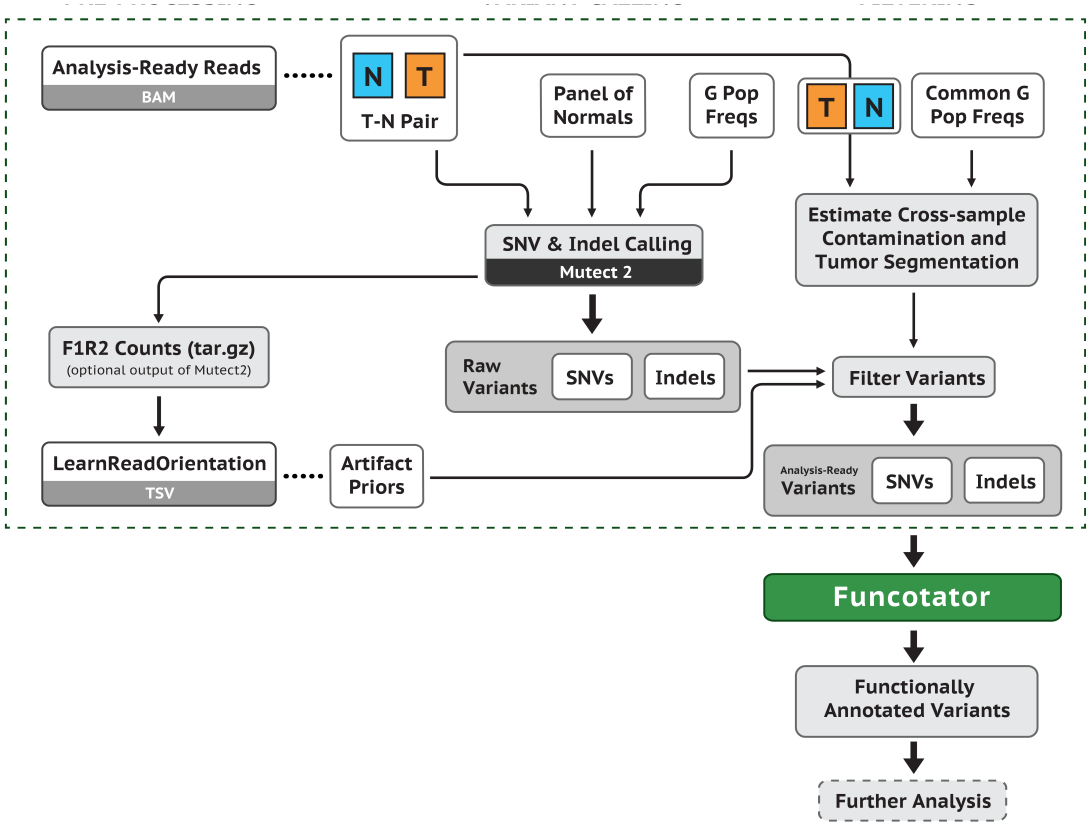


Figure: Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

MutEnricher?

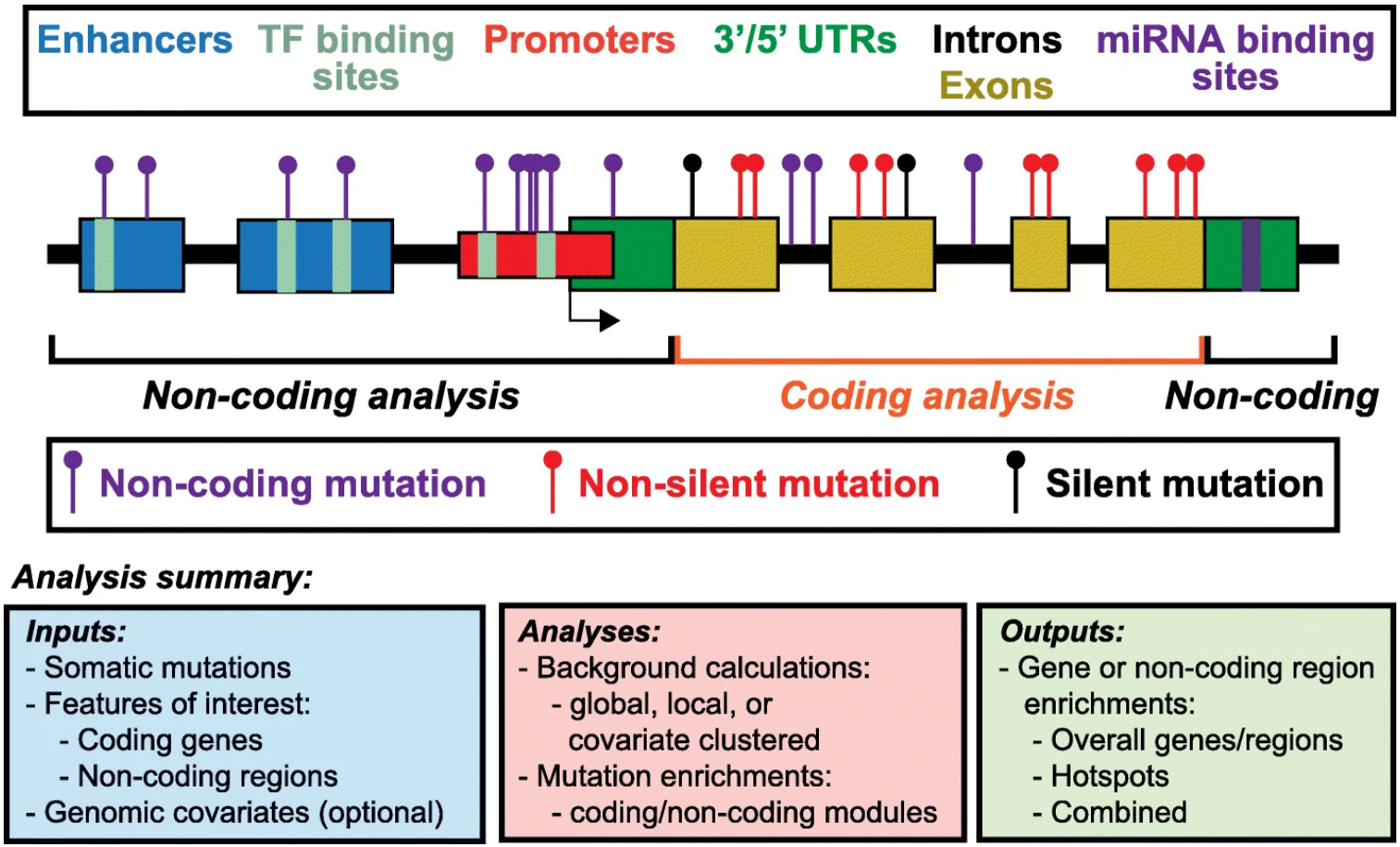


Figure: Schematic representation of MunEnricher's analysis procedures (Soltis et al., 2020)

Driver Gene Selection Strategy

COSMIC Cancer Gene Census (Tate John et al., 2018)

Gene $\in$ CGC Tier 1 set

Fisher FDR

Fisher FDR < 0.05

Fisher P-value

Fisher P-value < 0.05

Gene P-value

Gene P-value < 0.05

Somatic Variant in SQC

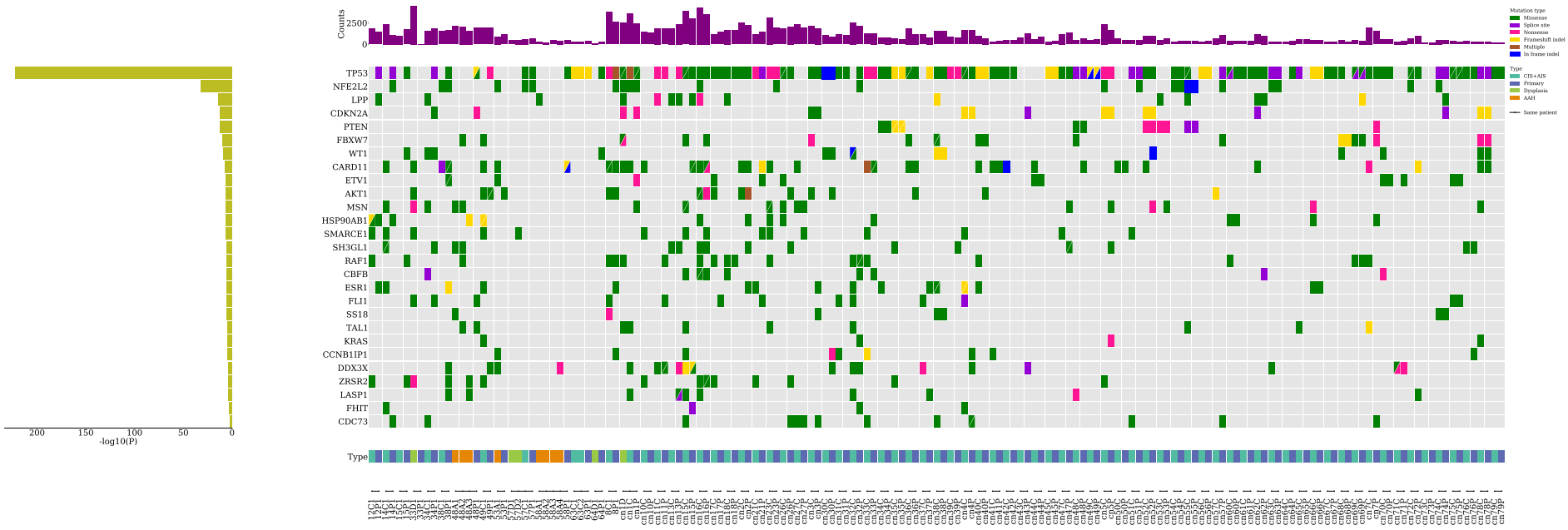


Figure: CoMut Plot with SQC Patients

Somatic Variant in ADC

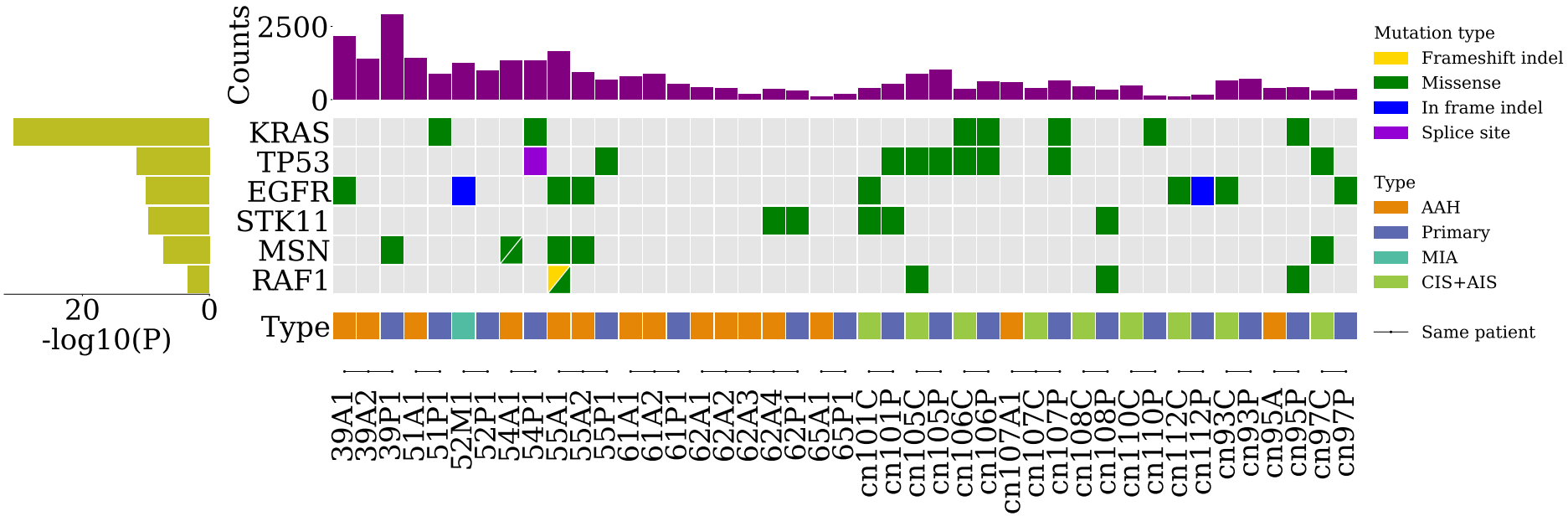


Figure: CoMut Plot with ADC Patients

Findings in SNVs Analysis

4. Results

4.4. VAF Analysis

VAF?

- Variant allele frequency
- VAF = Alternative allele read count/Total read count
- To find tumor evolution

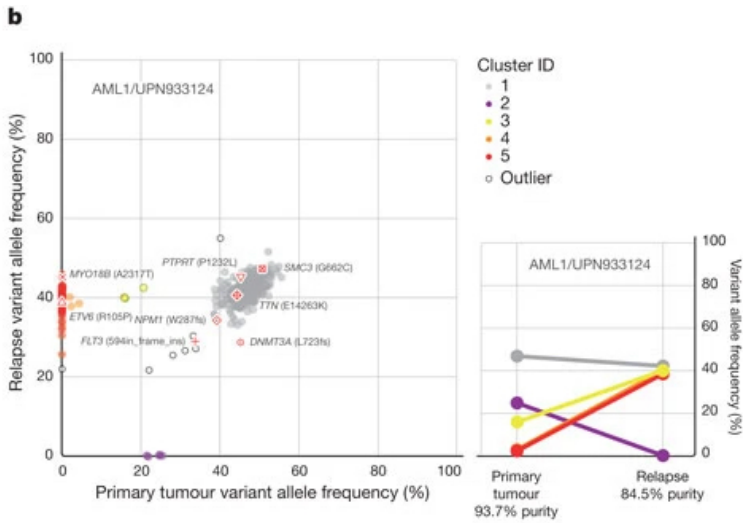


Figure: VAF distribution of validated mutations (Ding et al., 2012)

VAF Plots I

PyClone?

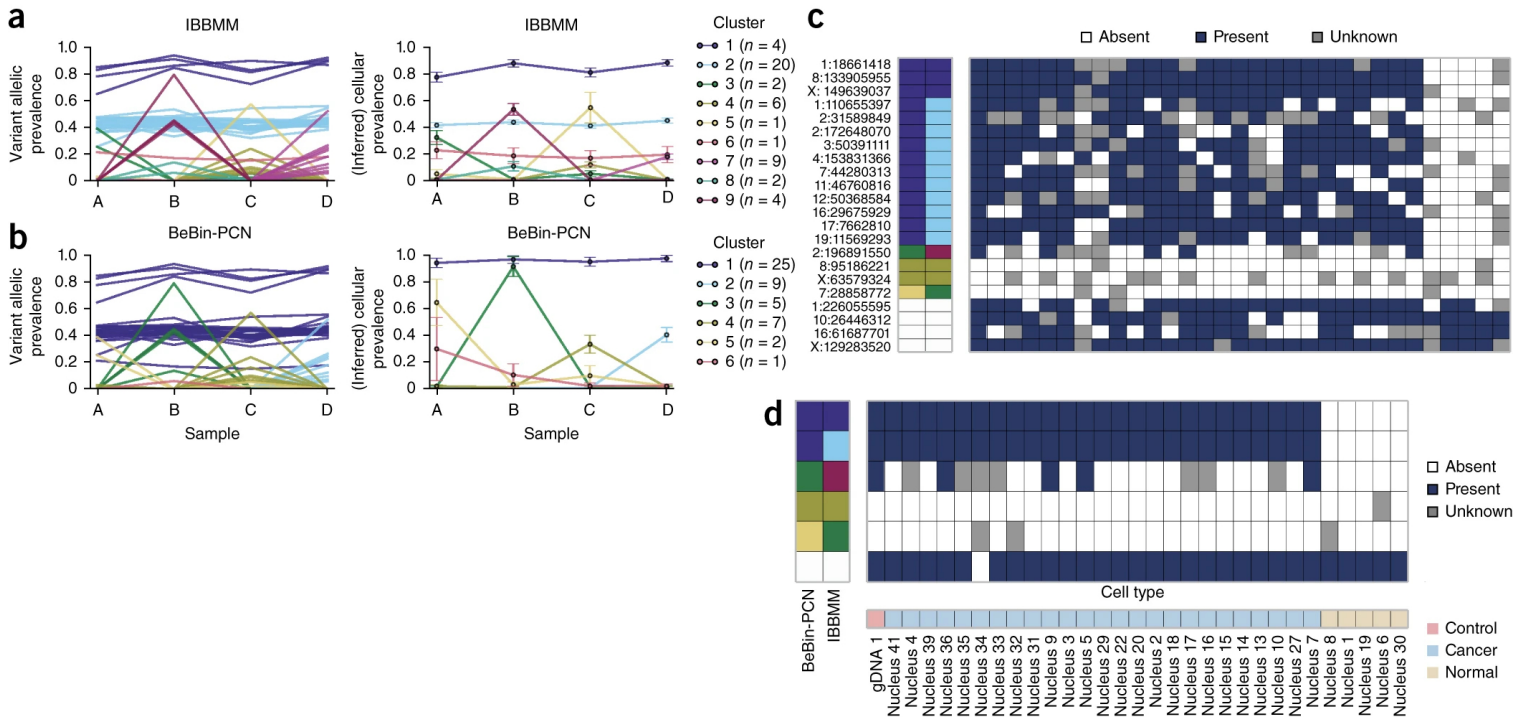


Figure: Analysis of multiple samples by PyClone (Roth et al., 2014)

PyClone Plots I

Findings in VAF Analysis

4. Results

4.5. Tumor Evolution Trajectories Analysis

Revolver?

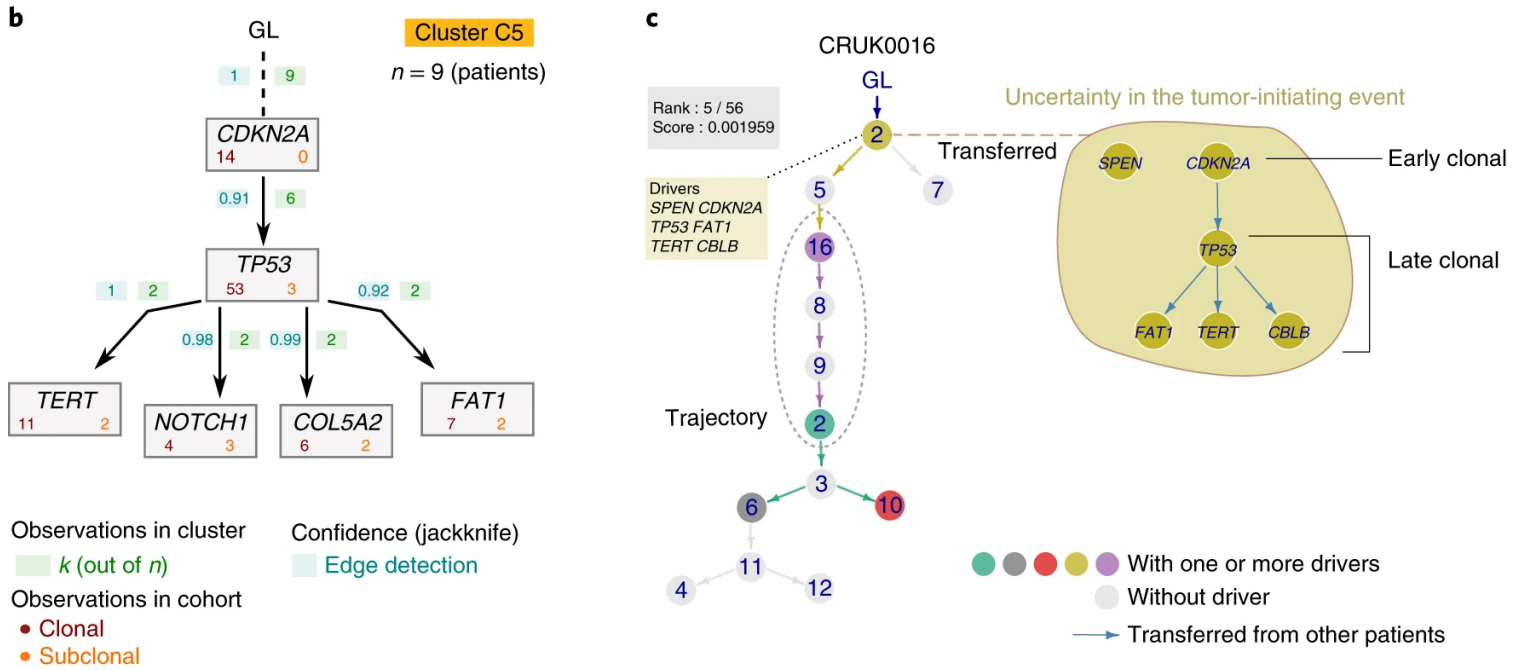


Figure: Repeated Evolutionary Trajectories (Caravagna et al., 2018)

Findings in Tumor Evolution Trajectories Analysis

4. Results

4.6. Differences in Gene Expression Levels

RSEM?

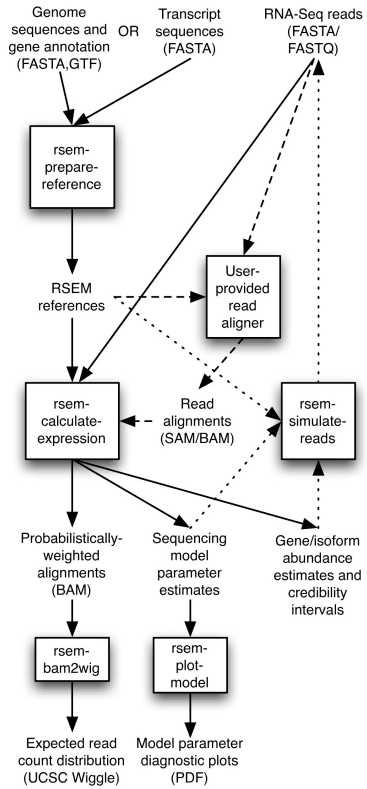


Figure: RSEM workflow (Li & Dewey, 2011)

DESeq2?

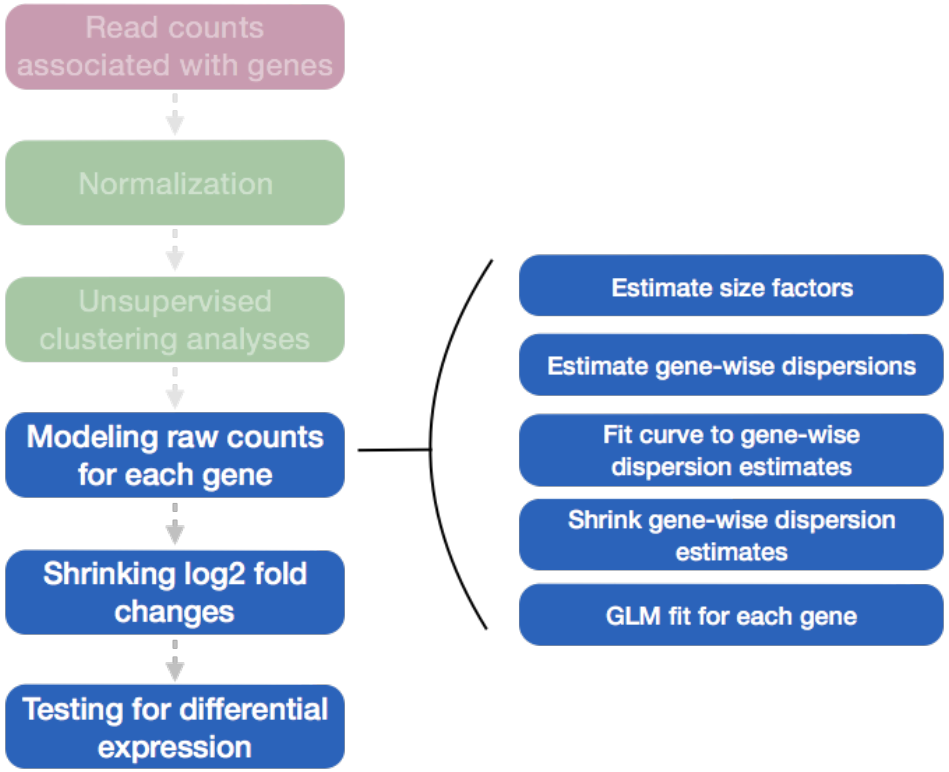


Figure: DESeq2 workflow (Love, Huber, & Anders, 2014)

DEG Selection Strategy

DEG: differentially expressed genes

Fold Change

$\log_2(\text{Fold Change}) > 1 \vee \log_2(\text{Fold Change}) < -1$

P-value

P-value < 0.05

Adjusted P-value

Padj < 0.05

Enrichr?

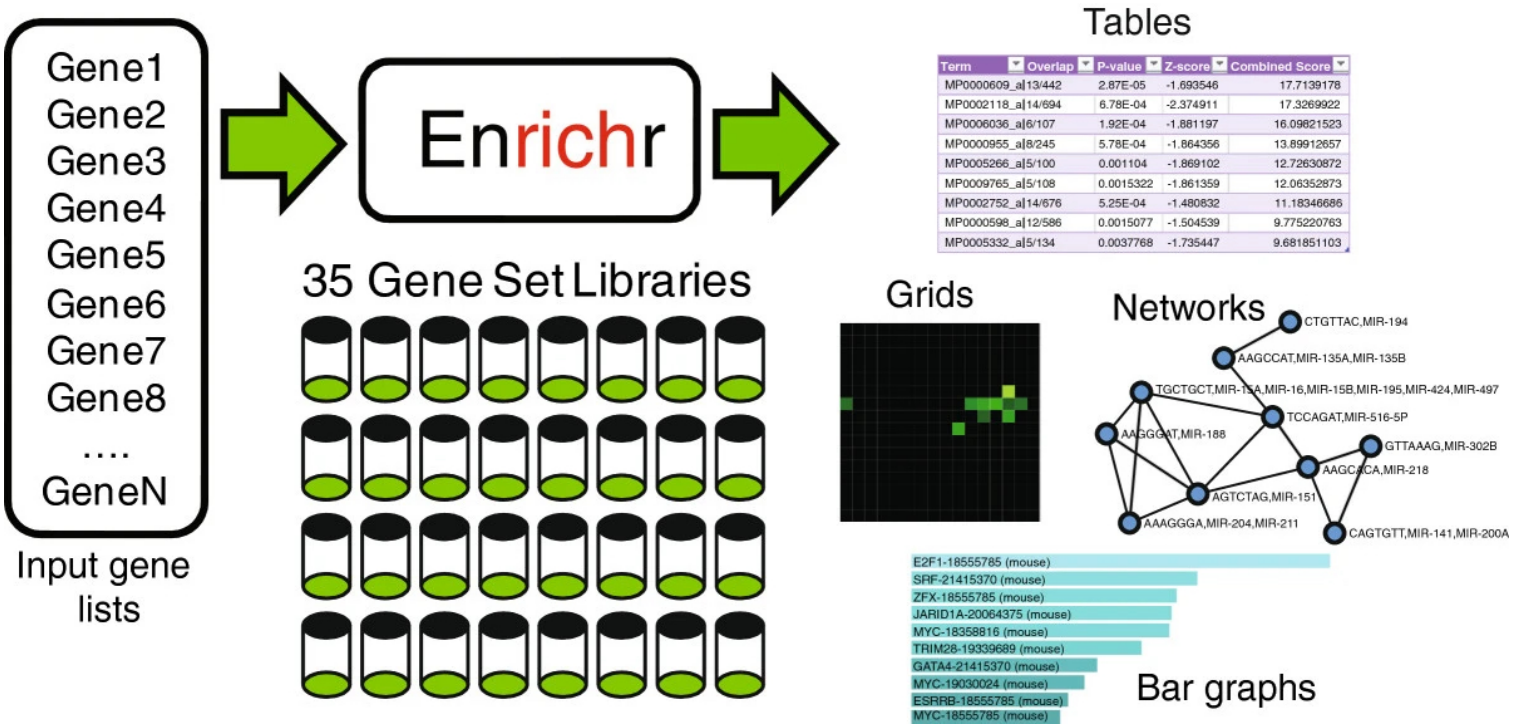


Figure: Enrichr workflow (Chen et al., 2013; Kuleshov et al., 2016)

Gene-set Library

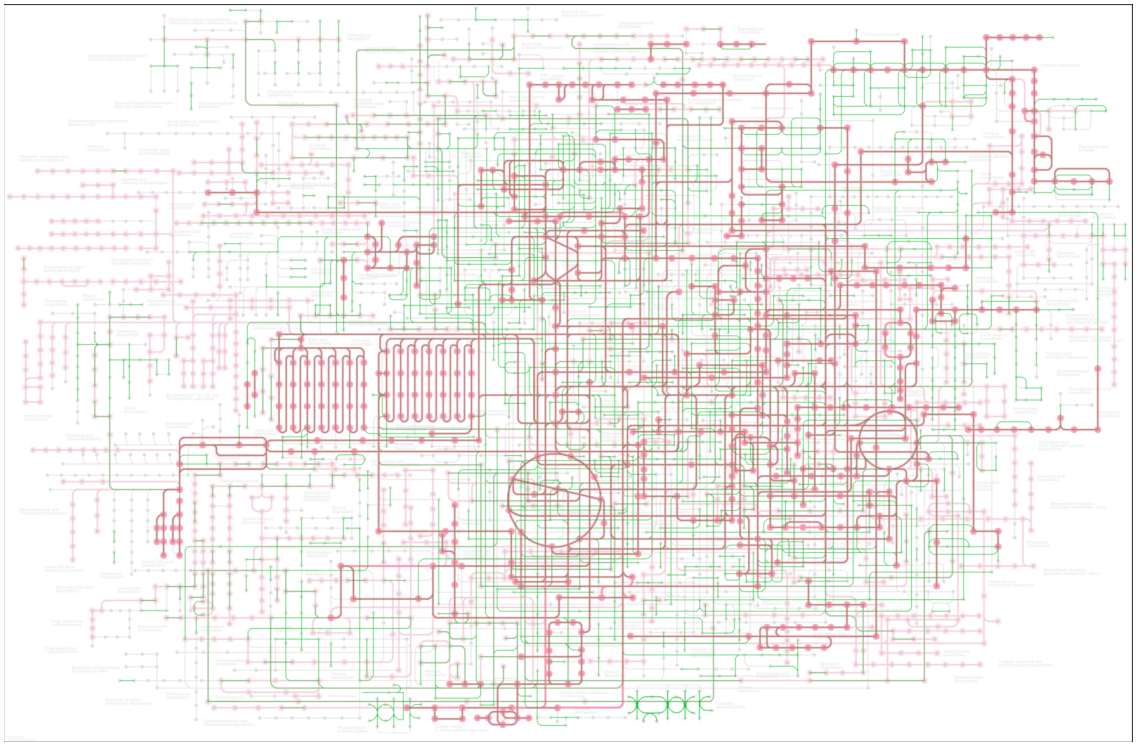


Figure: The global map of metabolic pathways by KEGG (Kanehisa et al., 2021)

KEGG

KEGG 2021 Human

WTS Data Composition I

Table: Number of WTS samples

Cancer Subtype	Stage	Number of Samples
SQC (n=89)	Normal	17
	Dysplasia	2
	CIS	33
	Primary	35
ADC (n=30)	Normal	12
	AAH	1
	AIS	9
	MIA	0
	Primary	8

WTS Data Composition II

Table: Number of WTS SQC samples

Recurrence?	Stage	Number of Samples
Recurrence (n=13)	Normal	1
	Dysplasia	1
	CIS	5
	Primary	6
Non-recurrence (n=74)	Normal	16
	Dysplasia	1
	CIS	28
	Primary	29

WTS Data Composition III

Table: Number of WTS ADC samples

Recurrence?	Stage	Number of samples
Recurrence (n=4)	Normal	1
	AAH	0
	AIS	2
	MIA	0
	Primary	1
Non-recurrence (n=26)	Normal	11
	AAH	1
	AIS	7
	MIA	0
	Primary	7

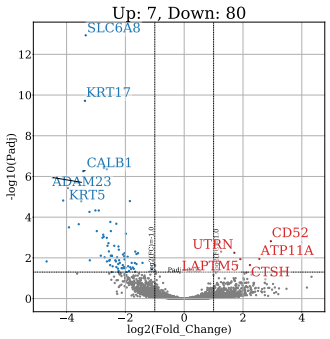
4. Results

4.6. Differences in Gene Expression Levels

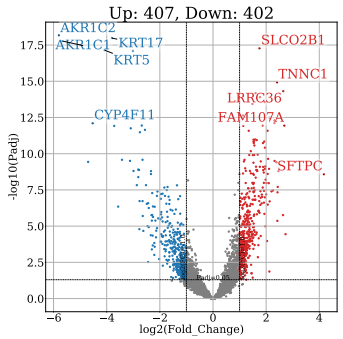
4.6.1. Comparing cancer stage

DEG Volcano Plots in SQC

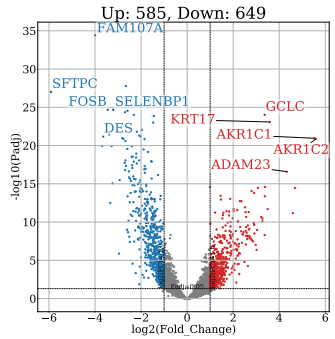
Normal → Dysplasia → CIS → Primary (SQC)



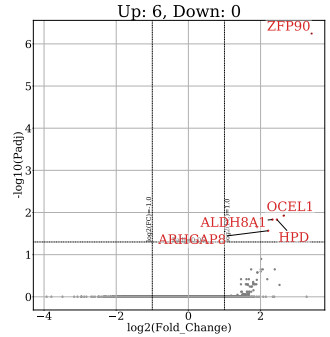
(a) Normal-Dysplasia



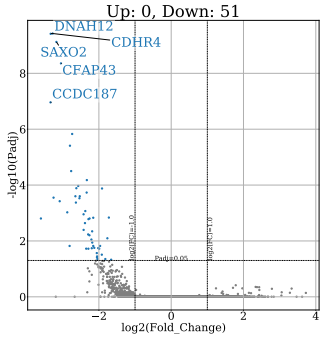
(b) Normal-CIS



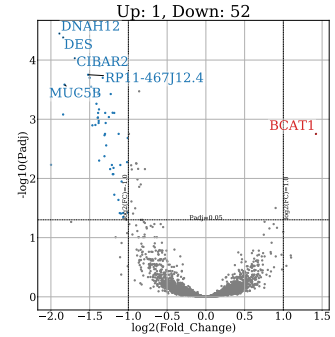
(c) Normal-Primary



(d) Dysplasia-CIS



(e) Dysplasia-Primary

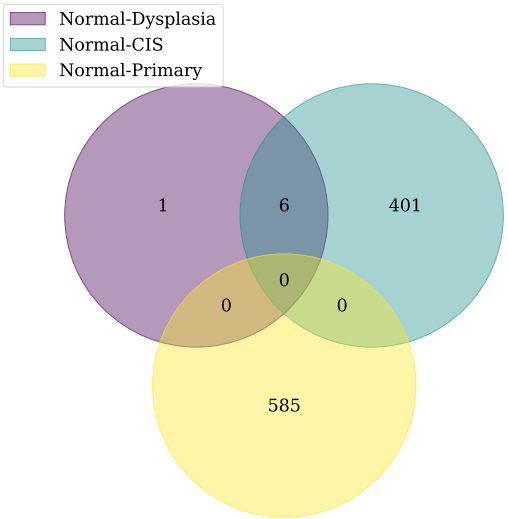


(f) CIS-Primary

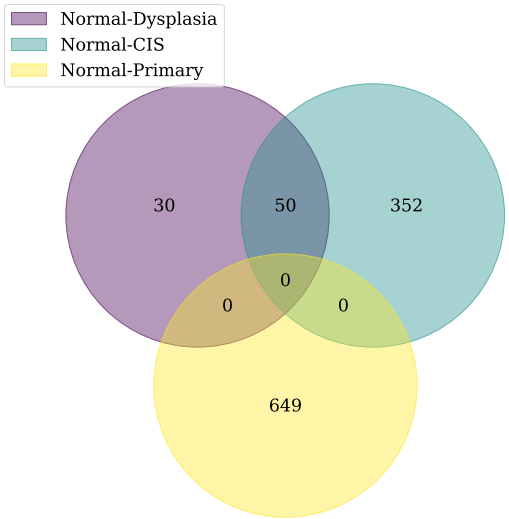
Figure: DEG Volcano Plots in SQC

DEG Venn Diagram in SQC

Normal → Dysplasia → CIS → Primary (SQC)



(a) Up-regulated



(b) Down-regulated

Figure: DEG Venn Diagram in SQC

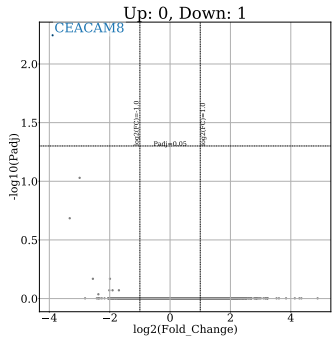
Enrichment test in SQC

Term name	Adjusted p-value
Glycolysis / Gluconeogenesis	1.88e-06
Metabolism of xenobiotics by cytochrome P450	5.19e-06
Glutathione metabolism	5.77e-06
Drug metabolism	5.77e-06
Central carbon metabolism in cancer	5.77e-06

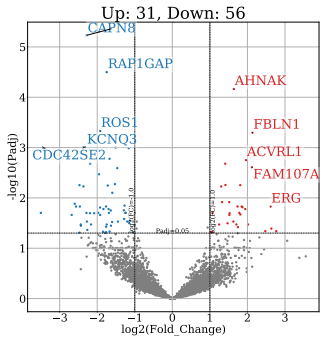
Term name	Adjusted p-value
Hematopoietic cell lineage	1.84e-10
Cell adhesion molecules	3.10e-07
Hypertrophic cardiomyopathy	3.10e-07
Malaria	3.10e-07
Viral myocarditis	2.04e-05

DEG Volcano Plots in ADC

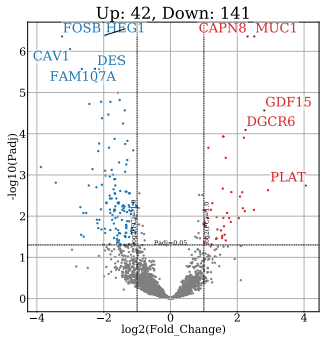
Normal → AAH → AIS → Primary (ADC)



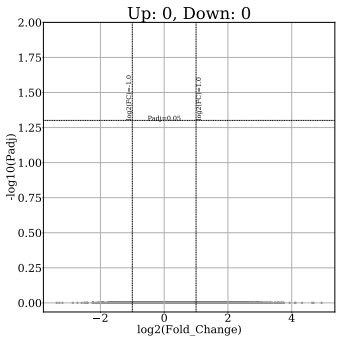
(a) Normal-AAH



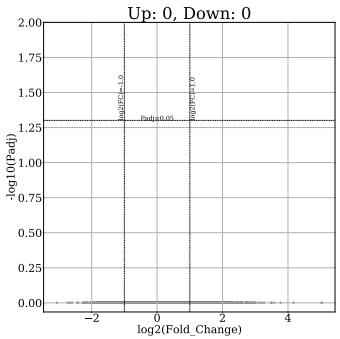
(b) Normal-AIS



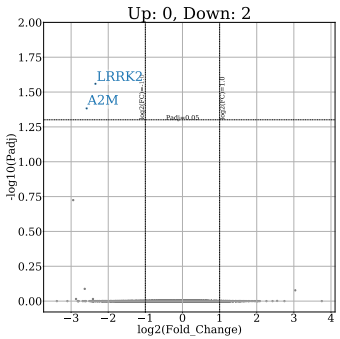
(c) Normal-Primary



(d) AAH-AIS



(e) AAH-Primary

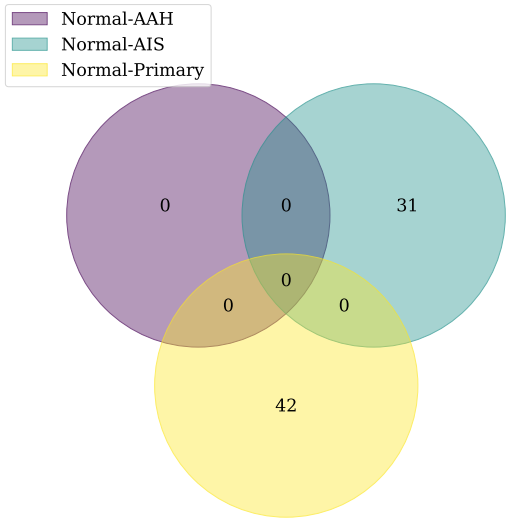


(f) AIS-Primary

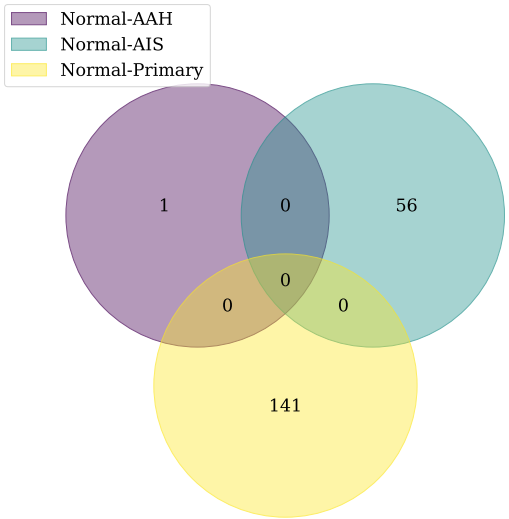
Figure: DEG Volcano Plots in ADC

DEG Venn Diagram in ADC

Normal $\rightarrow$ AAH $\rightarrow$ AIS $\rightarrow$ Primary (ADC)



(a) Up-regulated



(b) Down-regulated

Figure: DEG Venn Diagram in ADC

Enrichment test in ADC

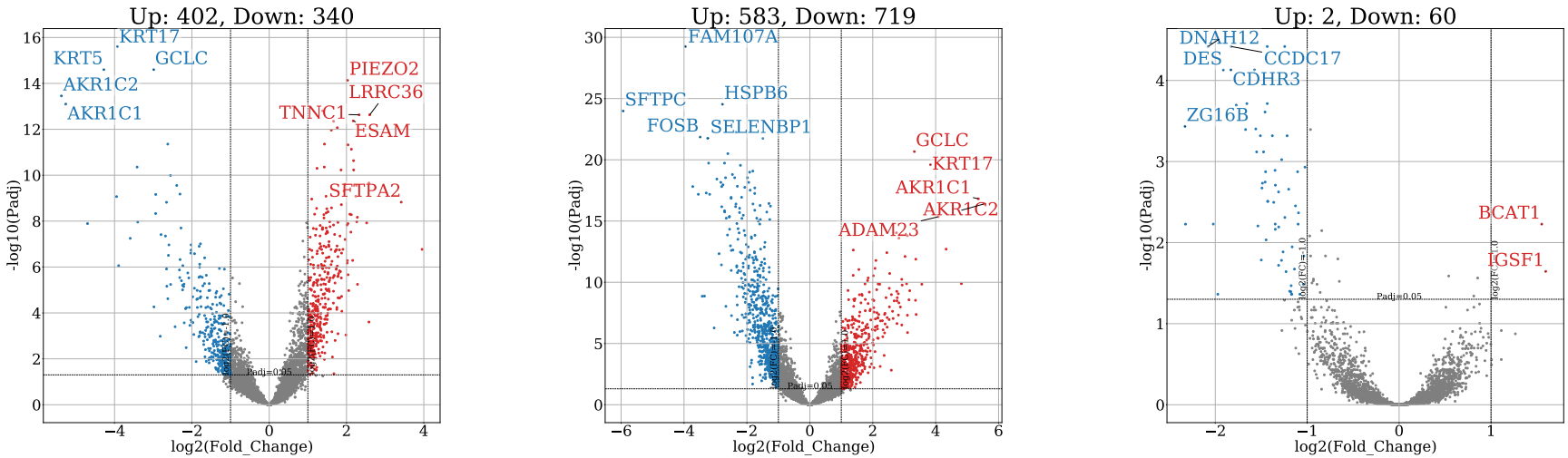
4. Results

4.6. Differences in Gene Expression Levels

4.6.2. Selecting Non-Recurrence & Comparing Stage

DEG Volcano Plots in SQC with Selecting Non-Recurrence

Normal → CIS → Primary

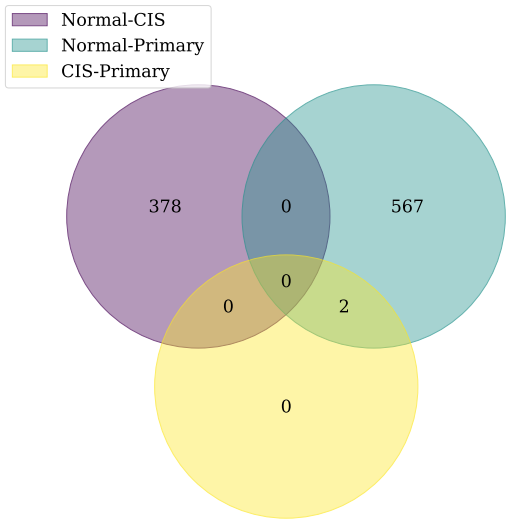


(a) Normal-CIS (b) Normal-Primary (c) CIS-Primary

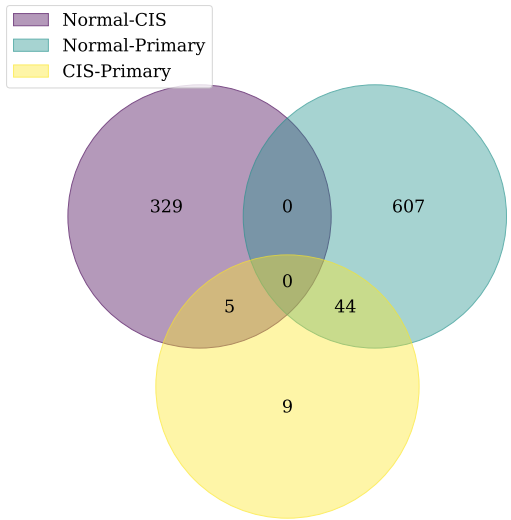
Figure: DEG Volcano Plots in SQC with Selecting Non-Recurrence

DEG Venn Diagram in SQC with Selecting Non-Recurrence

Normal $\rightarrow$ CIS $\rightarrow$ Primary



(a) Up-regulated



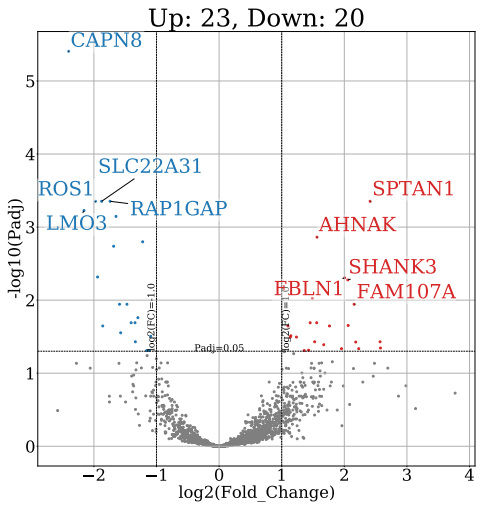
(b) Down-regulated

Figure: DEG Venn Diagram in SQC with Selecting Non-Recurrence

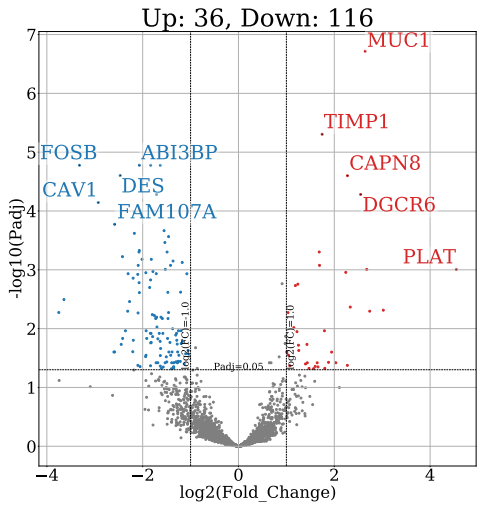
Enrichment test in SQC with Selecting Non-Recurrence

DEG Volcano Plots in ADC with Selecting Non-Recurrence

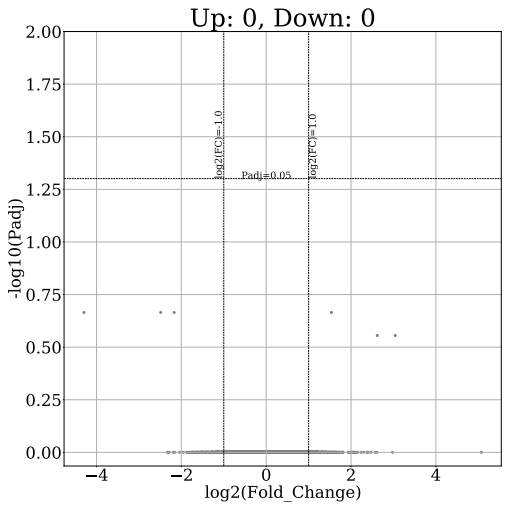
Normal → AIS → Primary



(a) Normal-AIS



(b) Normal-Primary

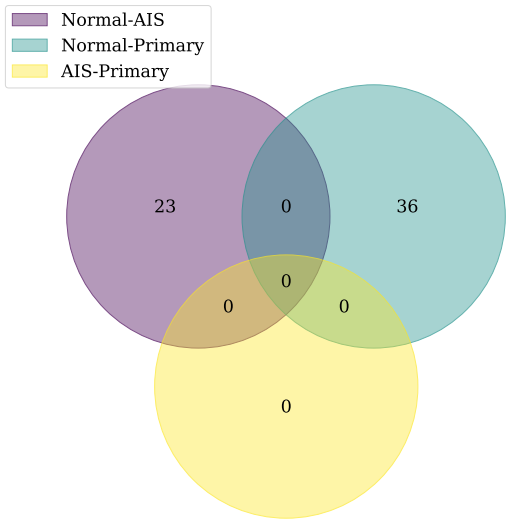


(c) AIS-Primary

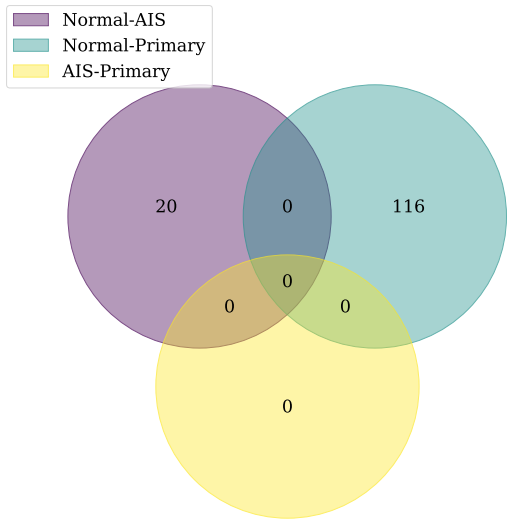
Figure: DEG Volcano Plots in ADC with Selecting Non-Recurrence

DEG Venn Diagram in ADC with Selecting Non-Recurrence

Normal $\rightarrow$ AIS $\rightarrow$ Primary



(a) Up-regulated



(b) Down-regulated

Figure: DEG Venn Diagram in ADC with Selecting Non-Recurrence

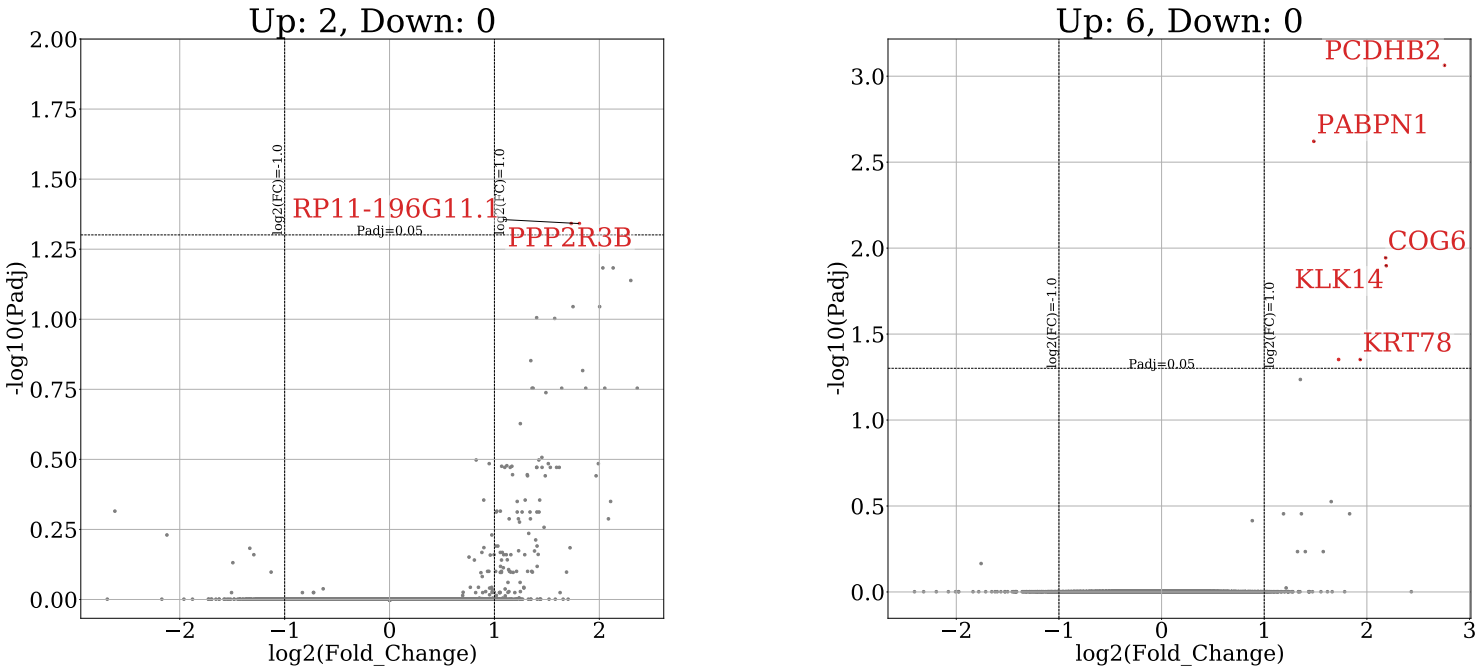
Enrichment test in ADC with Selecting Non-Recurrence

4. Results

4.6. Differences in Gene Expression Levels

4.6.3. Selecting Stage & Comparing Recurrence

DEG Volcano Plots in SQC with Selecting Stage



(a) CIS

(b) Primary

Figure: DEG Volcano Plots in SQC with Selecting Stage

Enrichment test in SQC with Selecting CIS

Enrichment test in SQC with Selecting Primary

Findings in DEG Analysis

4. Results

4.7. Bulk Cell Deconvolution

Single-cell data as Reference

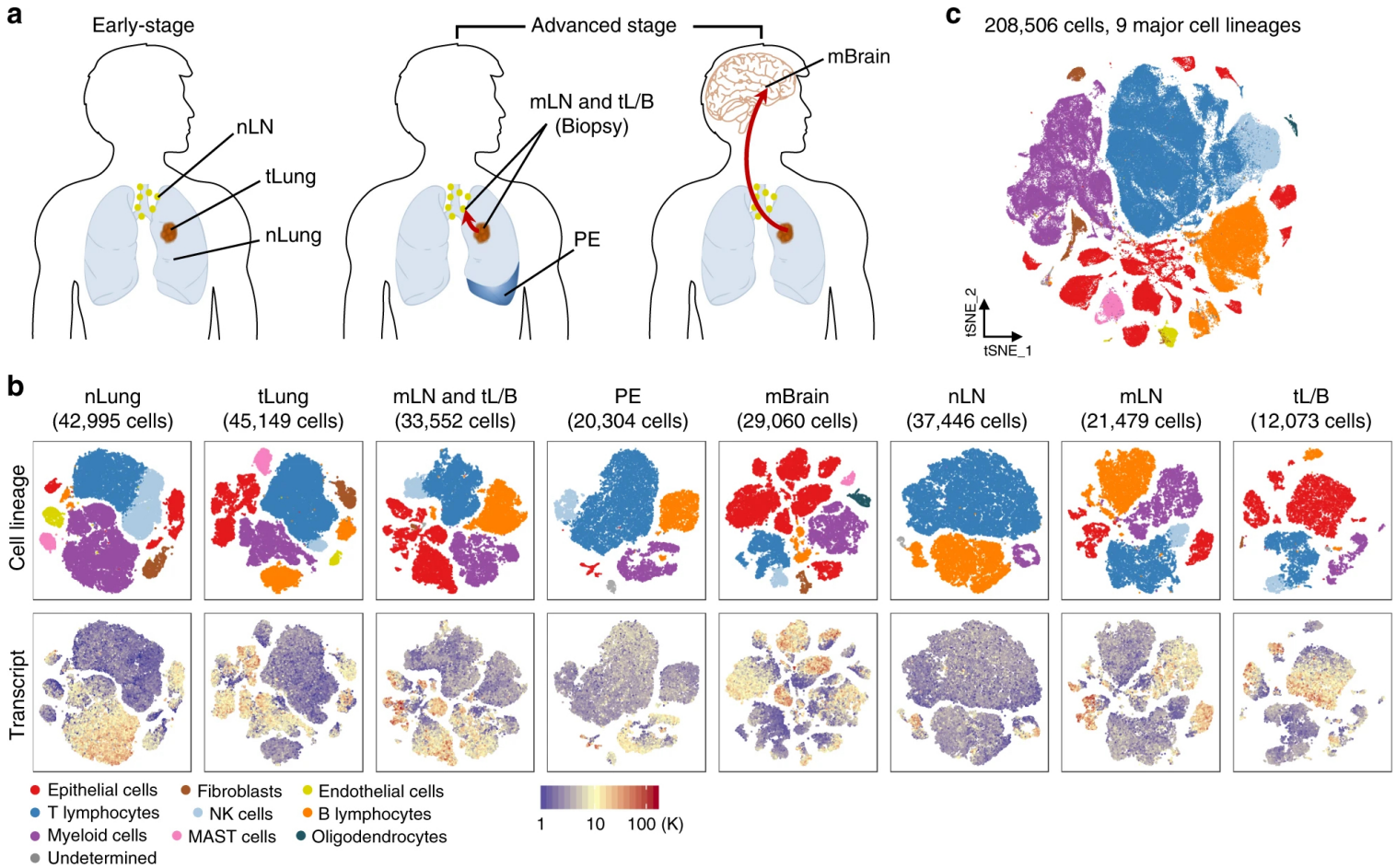


Figure: Comprehensive dissection and clustering of 208,506 single cells from LUAD patients (Kim et al., 2020)

MuSiC?

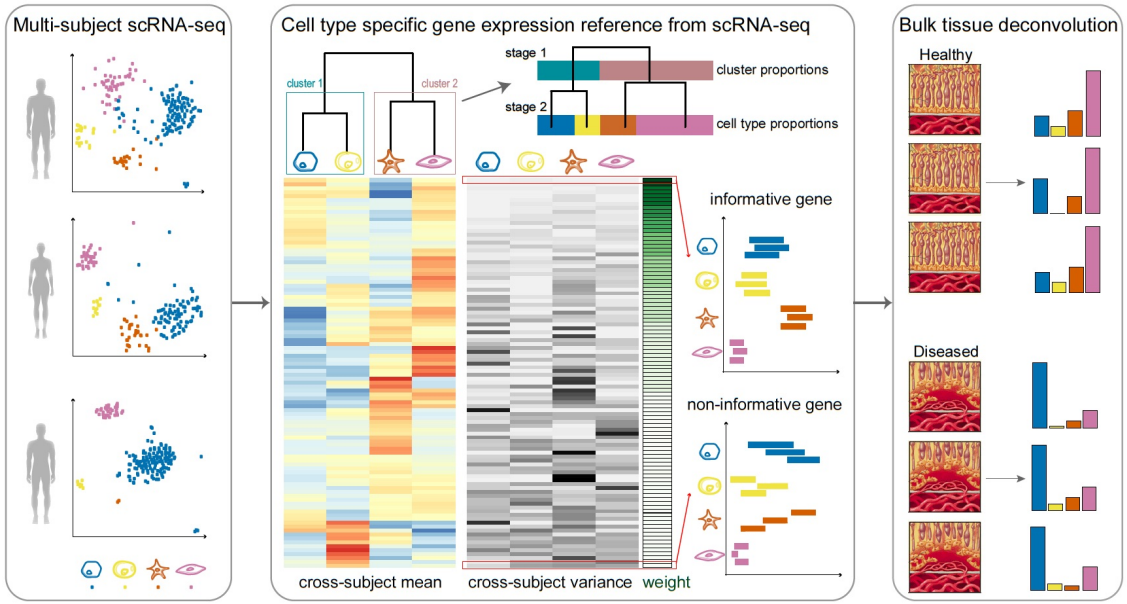


Figure: Workflow for MuSiC (X. Wang, Park, Susztak, Zhang, & Li, 2019)

Cluster Plot in SQC

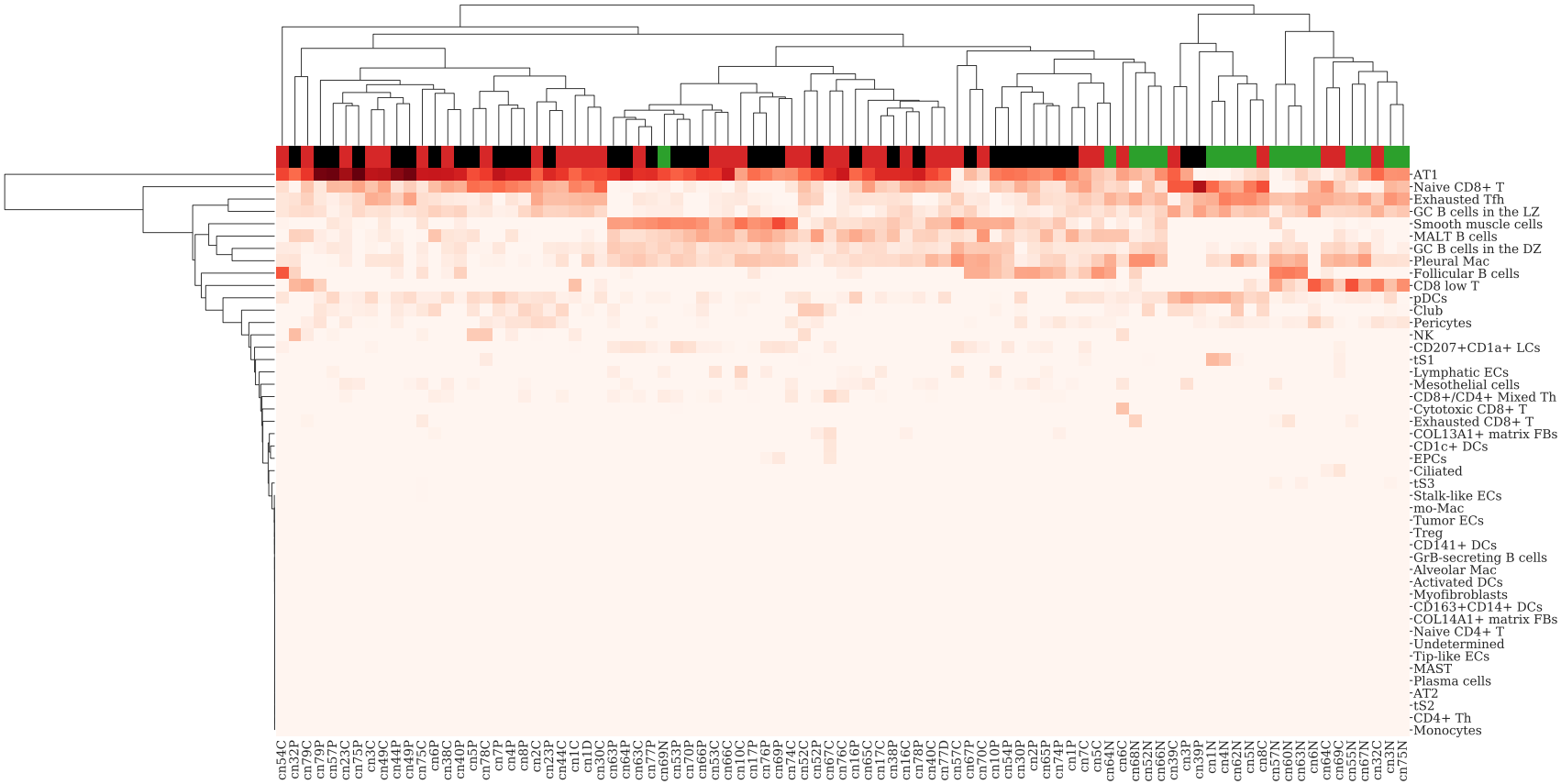


Figure: Cluster Plot in SQC

Violin Plots in SQC

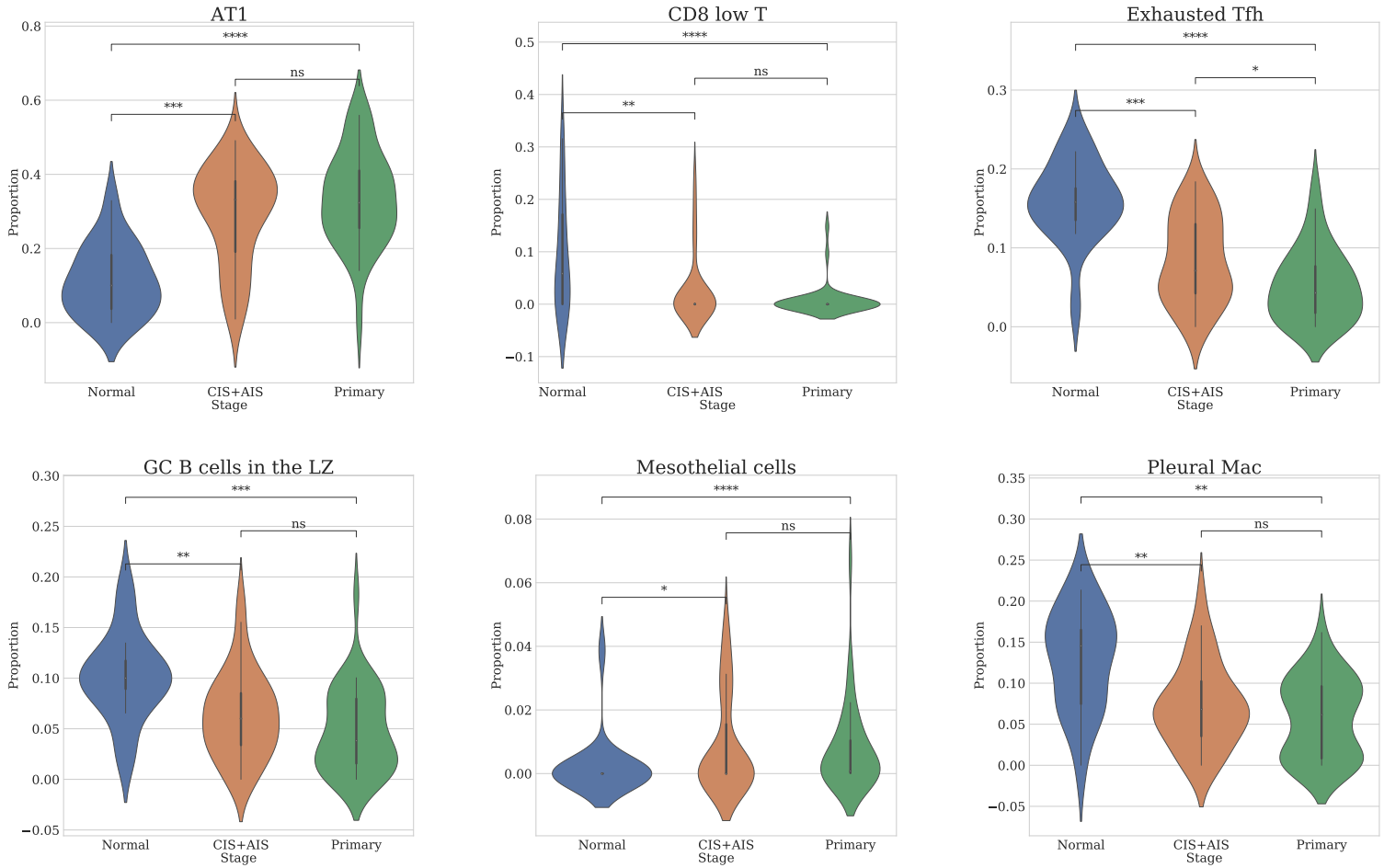


Figure: Violin Plots in SQC

Cluster Plot in ADC

Figure: Cluster Plot in ADC

Violin Plots in ADC

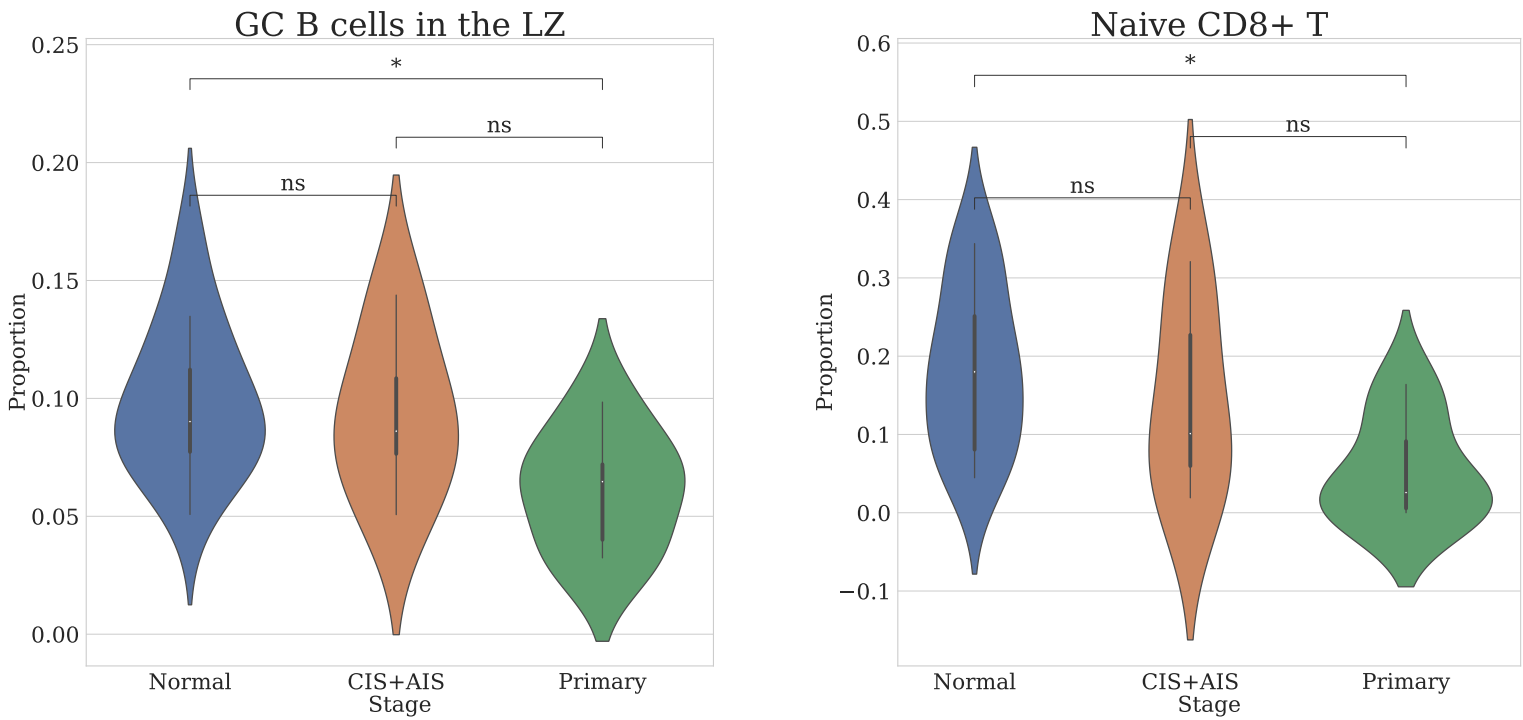


Figure: Violin Plots in ADC

Findings in Bulk Cell Deconvolution

4. Results

4.8. Discovery of Gene Fusion

Arriba?

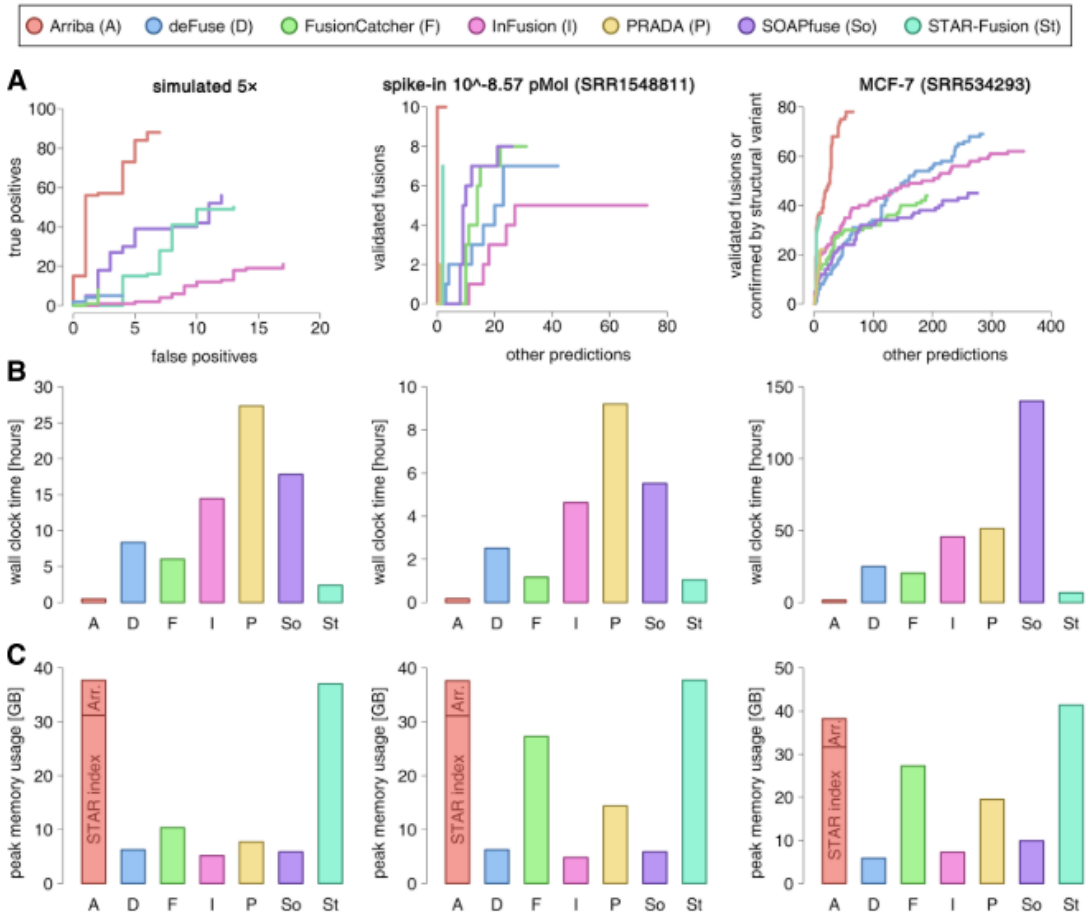


Figure: Benchmark of Arriba versus alternative methods (Uhrig et al., 2021)

Findings in Gene Fusion Discovery

5. Discussion

6. References

References I

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

Caravagna, G., Giarratano, Y., Ramazzotti, D., Tomlinson, I., Graham, T. A., Sanguinetti, G., & Sottoriva, A. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature methods*, *15*(9), 707–714.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., ... Ma'ayan, A. (2013). Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, *14*(1), 1–14.

Collins, L. G., Haines, C., Perkel, R., & Enck, R. E. (2007). Lung cancer: diagnosis and management. *American family physician*, *75*(1), 56–63.

References II

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.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., ... others (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, *481*(7382), 506–510.

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.

References III

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.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). Kegg: integrating viruses and cellular organisms. *Nucleic acids research*, *49*(D1), D545–D551.

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.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., ... others (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90–W97.

References IV

Li, B., & Dewey, C. N. (2011). Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, *12*(1), 1–16.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, *15*(12), 1–21.

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

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., ... Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, *11*(4), 396–398.

Soltis, A. R., Dalgard, C. L., Pollard, H. B., & Wilkerson, M. D. (2020). Mutenricher: a flexible toolset for somatic mutation enrichment analysis of tumor whole genomes. *BMC bioinformatics*, *21*(1), 1–8.

References V

Tate John, G., Sally, B., Jubb Harry, C., Zbyslaw, S., Beare David, M., Nidhi, B., ... Elisabeth, D. (2018). Stefancsik ray, thompson sam l, wang shicai, ward sari, campbell peter j, forbes simon a. cosmic: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, *47*(D1), D941–D947.

Travis, W. D. (2002). Pathology of lung cancer. *Clinics in chest medicine*, *23*(1), 65–81.

Uhrig, S., Ellermann, J., Walther, T., Burkhardt, P., Fröhlich, M., Hutter, B., ... others (2021). Accurate and efficient detection of gene fusions from rna sequencing data. *Genome research*, *31*(3), 448–460.

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.

References VI

Vincent, R. G., Pickren, J. W., Lane, W. W., Bross, I., Takita, H., Houten, L., ... Rzepka, T. (1977). The changing histopathology of lung cancer. a review of 1682 cases. *Cancer*, *39*(4), 1647–1655.

Wang, B.-Y., Huang, J.-Y., Chen, H.-C., Lin, C.-H., Lin, S.-H., Hung, W.-H., & Cheng, Y.-F. (2020). The comparison between adenocarcinoma and squamous cell carcinoma in lung cancer patients. *Journal of cancer research and clinical oncology*, *146*(1), 43–52.

Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, *10*(1), 1–9.