

# Lung Precancer Study

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

Department of Biomedical Engineering  
Ulsan National Institute of Science and Technology

*jwlee230@unist.ac.kr*

2021-07-09

# Overview

1 Introduction

2 Materials

3 Methods

4 Results

5 Discussion

6 References

# Introduction

# Introduction

## Lung Cancer

# Lung Cancer? I

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)

- Thyroid: 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

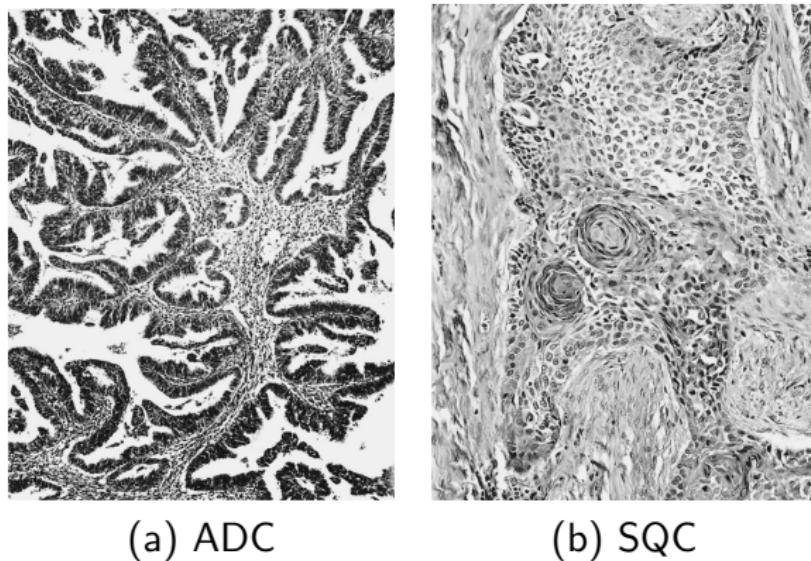
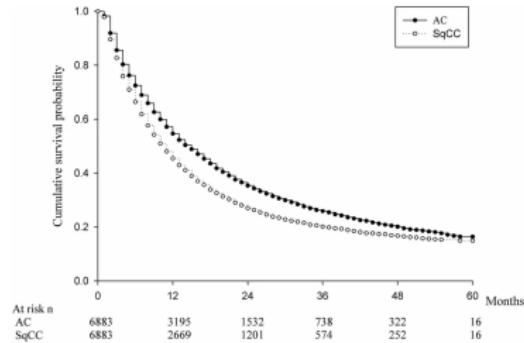
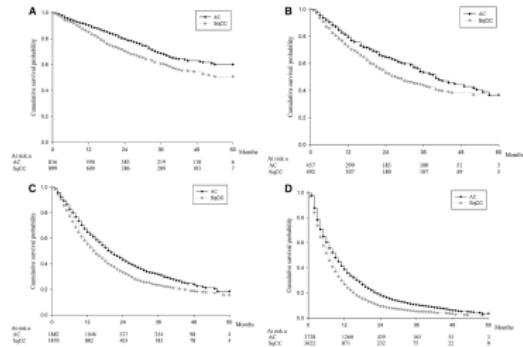


Figure: ADC and SQC histology in Lung cancer (Travis, 2002)

# ADC vs. SQC II



(a) All patients



(b) By cancer stages

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

## Findings

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

## Introduction

## 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

# Materials

# Lung Cancer Data

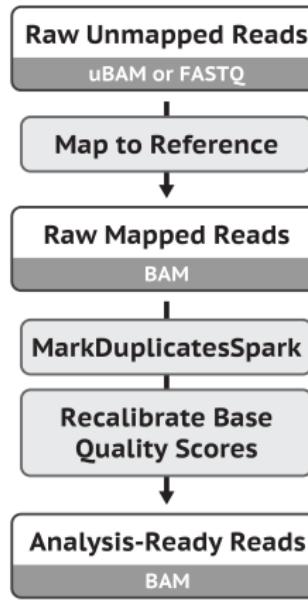
- WES (n=289) + Transcriptome (n=166)
- Normal + {Primary, CIS + AIS, AAH, Dysplasia, MIA}
  - Carcinoma in situ
  - Adenocarcinoma in situ
  - Atypical adenomatous hyperplasia
  - Dysplasia
  - Minimally invasive adenocarcinoma
- Squamous cell carcinoma (SQC) & Adenocarcinoma (ADC)
  - ① Normal → Dysplasia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

# Methods

## Methods

## Workflows

# Data pre-processing for variant discovery



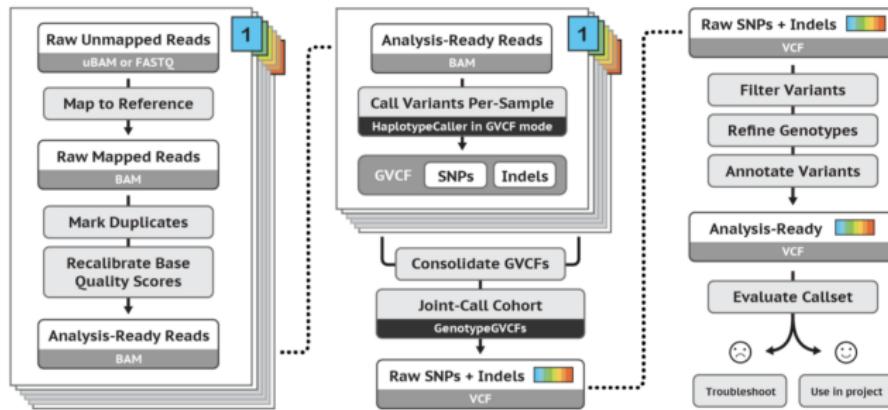
**Figure:** Data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

# Somatic short variant discovery



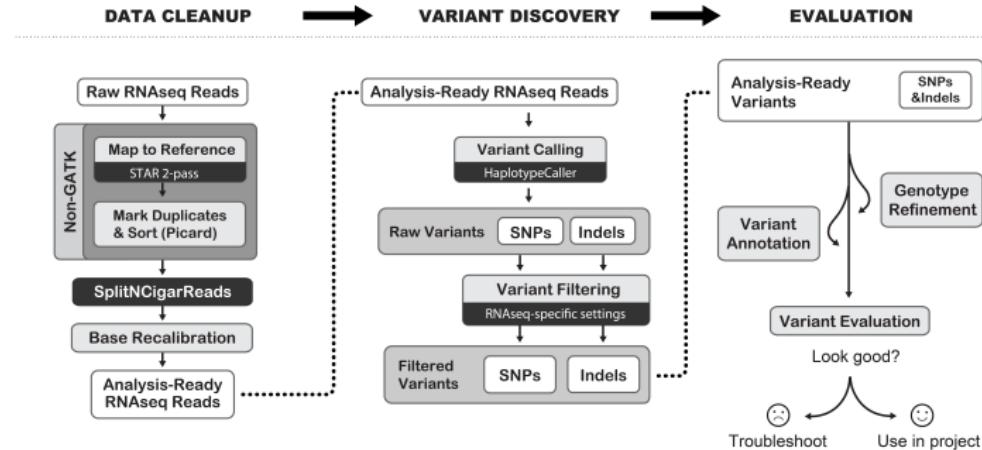
**Figure:** Somatic short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

# 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)

## Methods

## Miscellaneous

# Used Bioinformatics Tools

- BWA (H. Li & Durbin, 2009; H. Li, 2013)
- STAR (Dobin et al., 2013)
- Bowtie2 (Langmead & Salzberg, 2012)
- Samtools (H. Li et al., 2009)
- GATK (Van der Auwera et al., 2013; DePristo et al., 2011)
- Picard (*Picard toolkit*, 2019)
- VCF2MAF (Kandoth et al., 2018)
- BCFtools (Danecek et al., 2021)
- VEP (McLaren et al., 2016)

# R Packages

- Sequenza (Favero et al., 2015)
- Copynumber (Nilsen, Liestol, & Lingjaerde, 2013; Nilsen et al., 2012)
- DESeq2 (Love, Huber, & Anders, 2014)

# Python Packages

- Pandas (pandas development team, 2020; Wes McKinney, 2010)
- Sequenza-utils (Favero et al., 2015)
- Matplotlib (Hunter, 2007)
- Seaborn (Waskom & the seaborn development team, 2020)
- CoMut (Crowdis, He, Reardon, & Van Allen, 2020)
- PyClone (Roth et al., 2014)
- Statannot

# Results

# Results

## Quality Checks with FastQC

# FastQC?

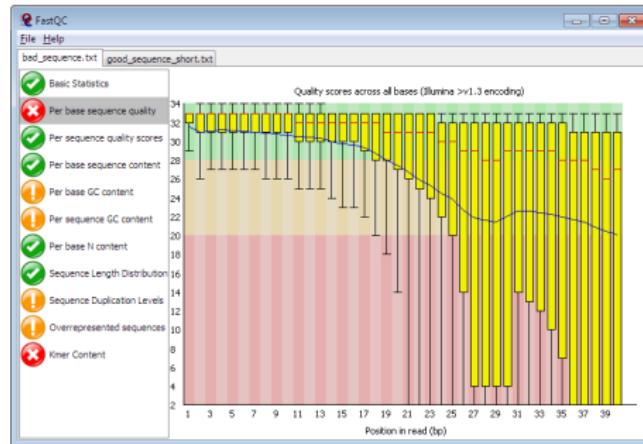


Figure: Example of FastQC Result (Andrews et al., 2012)

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

# FastQC on WES

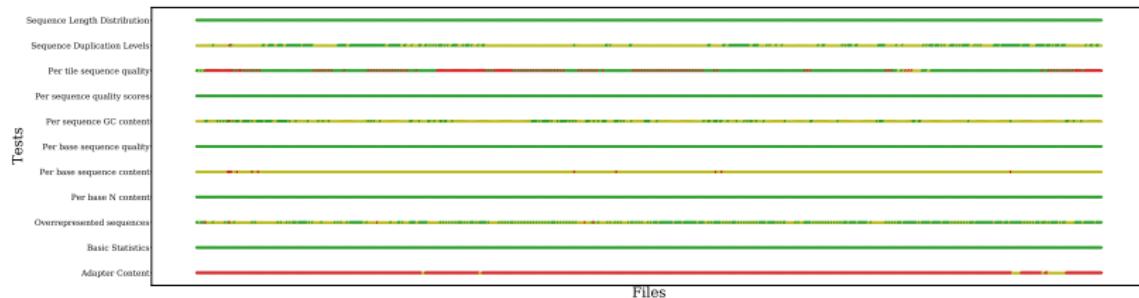
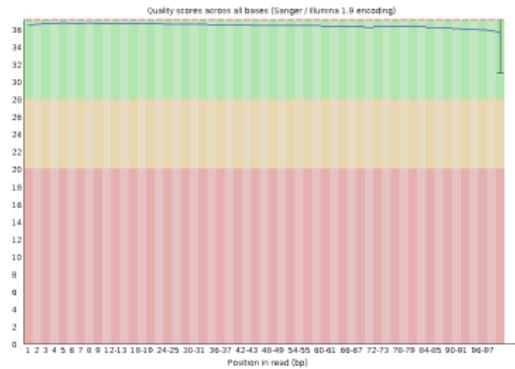


Figure: FastQC with WES data

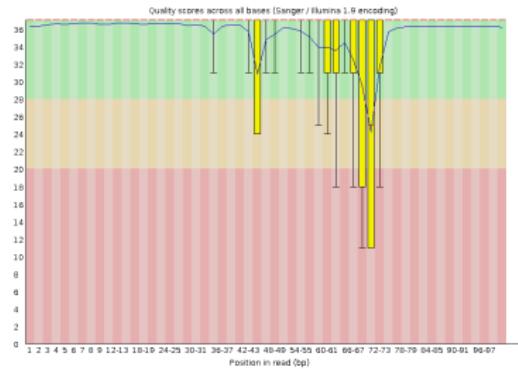
Failure on 33P1 sample

33P1 is excluded at further analysis.

# Failure on 33P1 I



(a) 33N



(b) 33P1

Figure: Per Base Sequence Quality Results

# Failure on 33P1 II

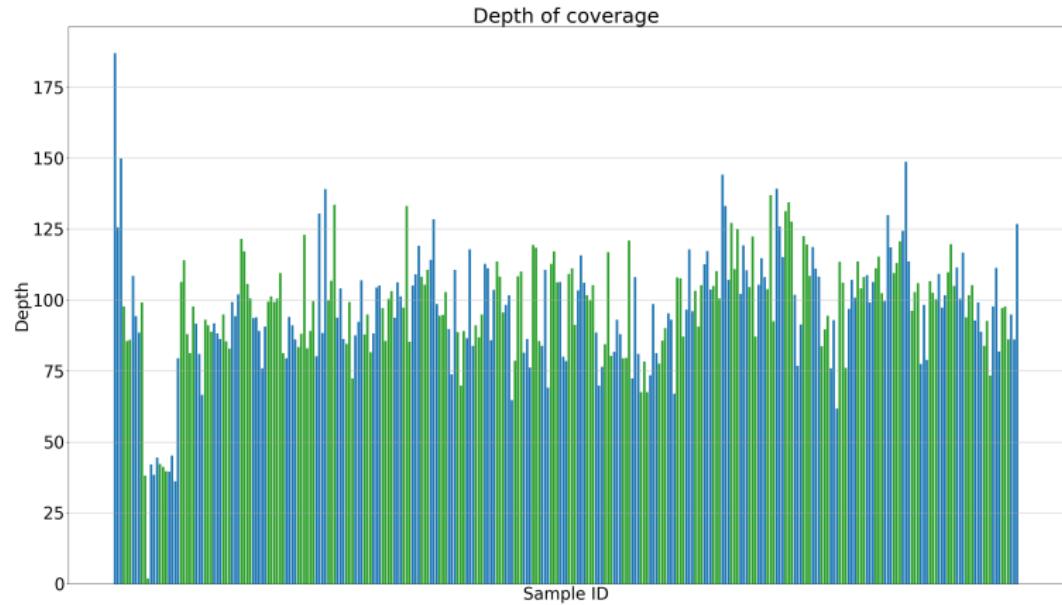


Figure: Coverage Depth Plot

# FastQC on WTS

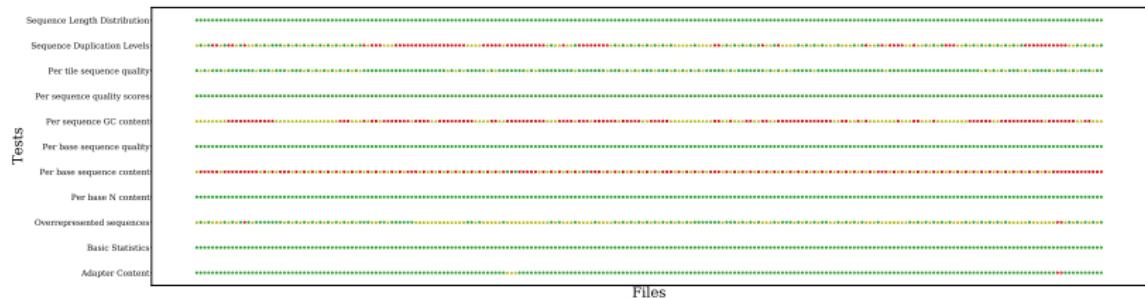


Figure: FastQC with WTS data

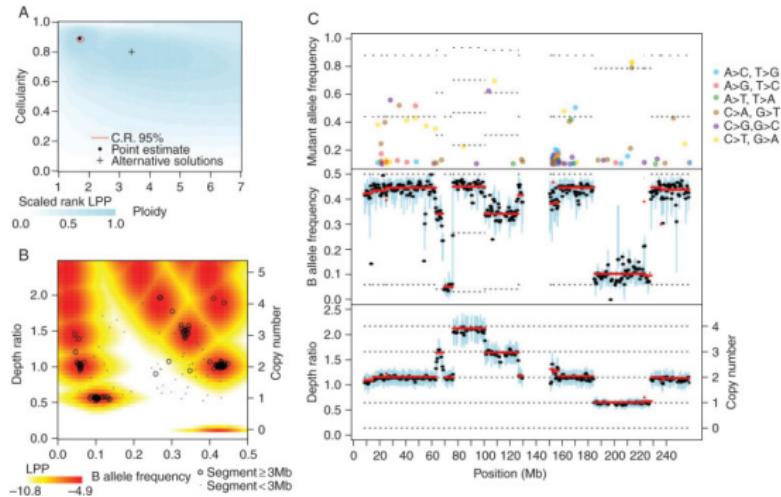
All sample are good to analysis

∴ No sample has more than 5 failures.

# Results

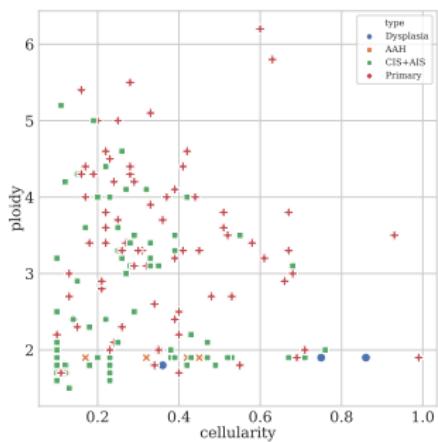
Copy Number Variations (CNVs) with Sequenza

# Sequenza?

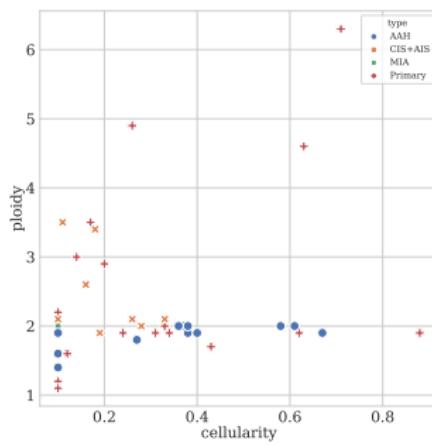


**Figure:** Representative Output of the Sequenza (Favero et al., 2015)

# Cellularity & Ploidy on WES



(a) SQC Samples



(b) ADC Samples

Figure: Cellularity and Ploidy from Sequenza

# Genome View on Patient #57

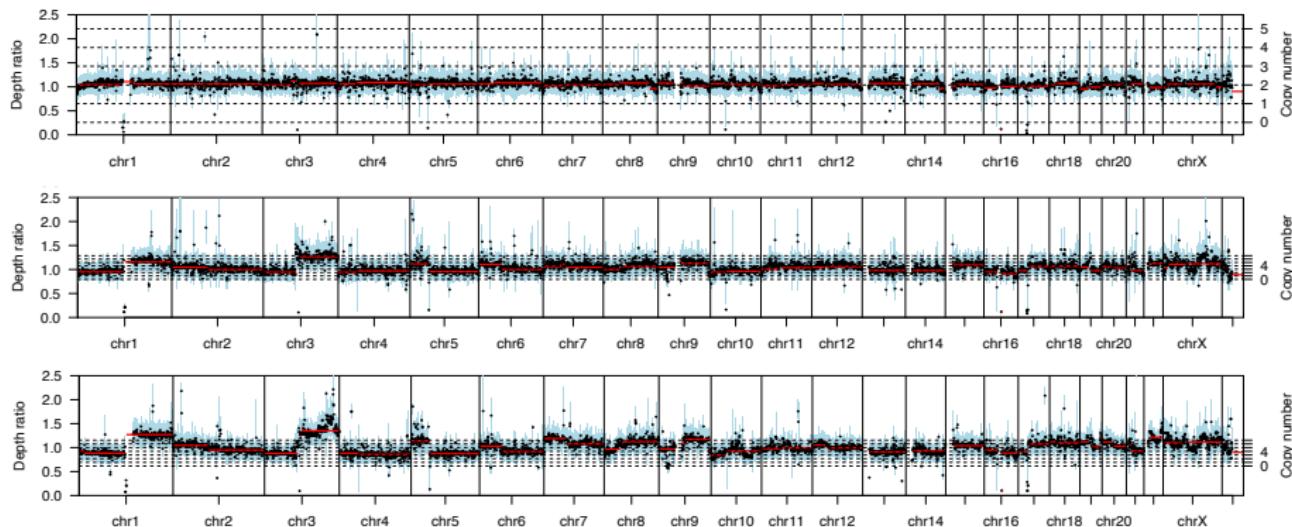


Figure: Dysplasia-CIS-Primary Tumor on Patient #57

# CNVs of SQC

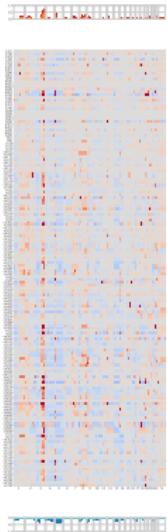
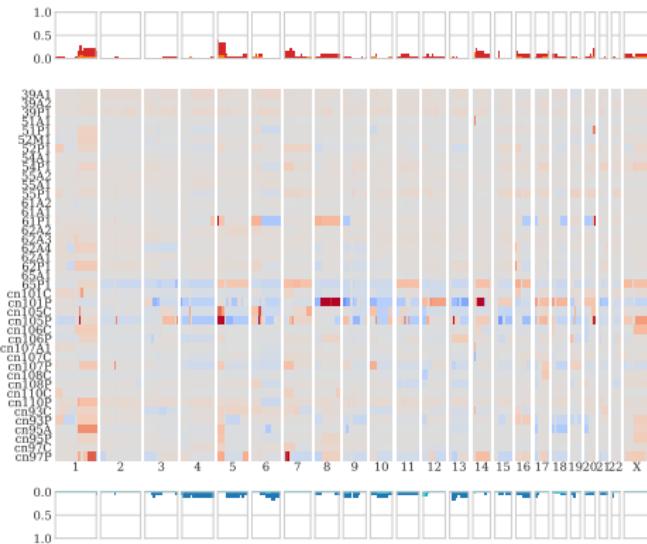


Figure: CNV Plot with SQC Patients

## CNVs of ADC



## Figure: CNV Plot with ADC Patients

# SQC vs. ADC

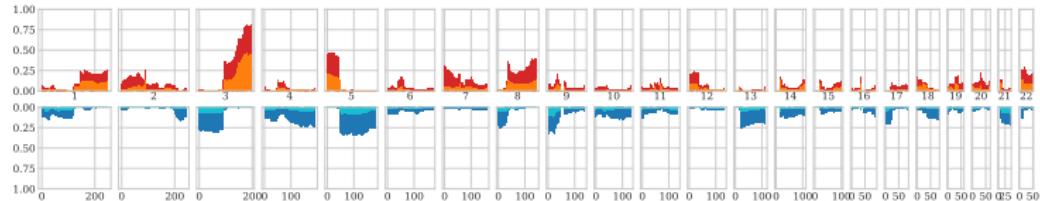


Figure: Simple CNV Plot with SQC Patients

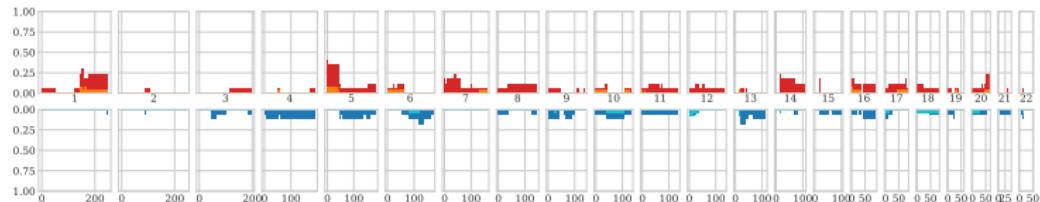


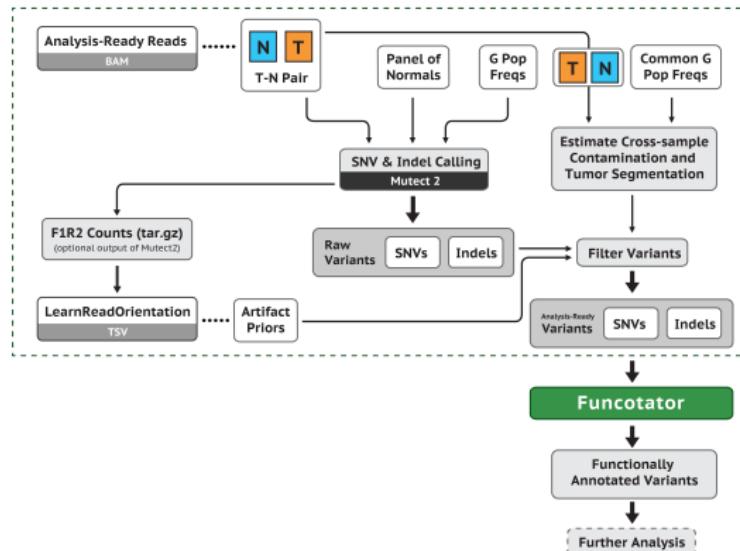
Figure: Simple CNV Plot with ADC Patients

# Findings in Sequenza

# Results

## SNVs Analysis

# Mutect2?



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

# Witer?

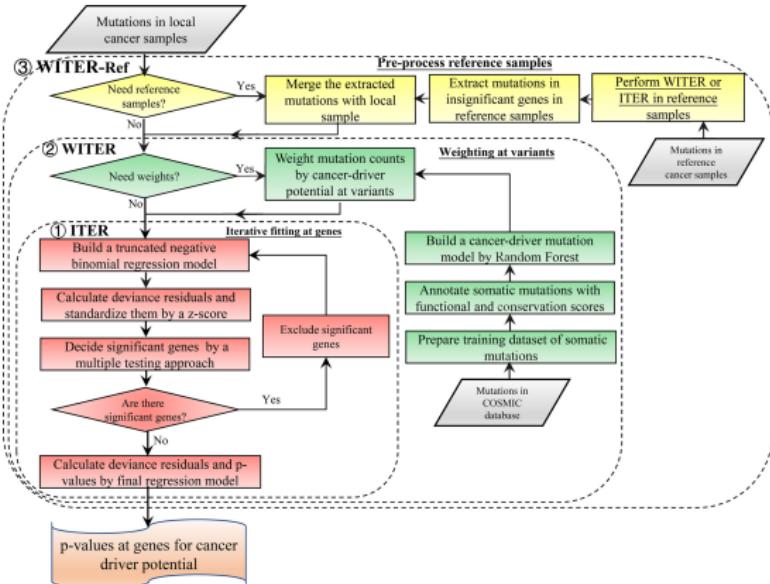


Figure: Witer diagram for detecting cancer-drive genes (Jiang et al., 2019)

# Somatic Variant with BWA in SQC

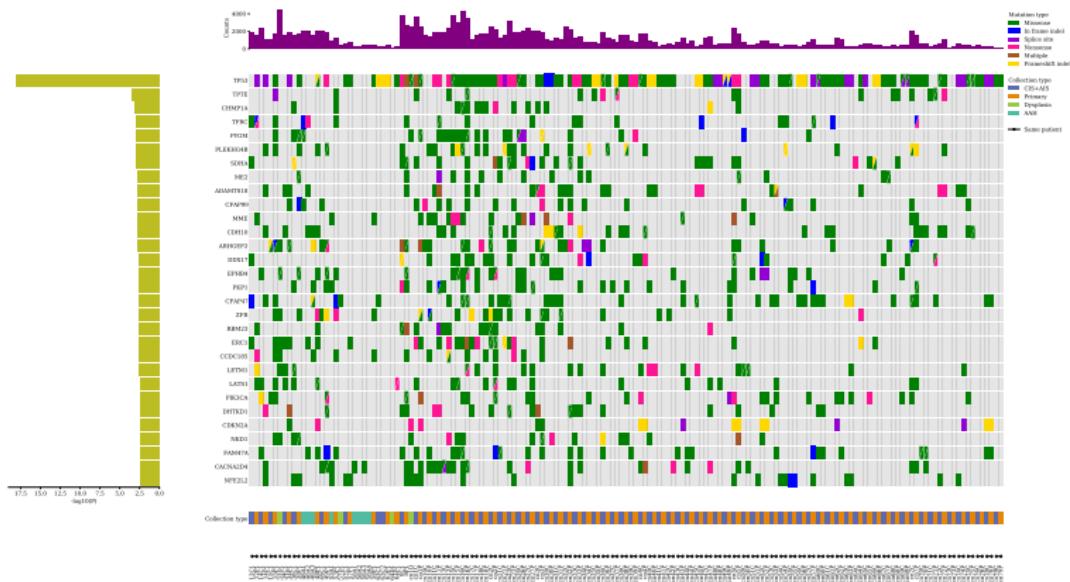
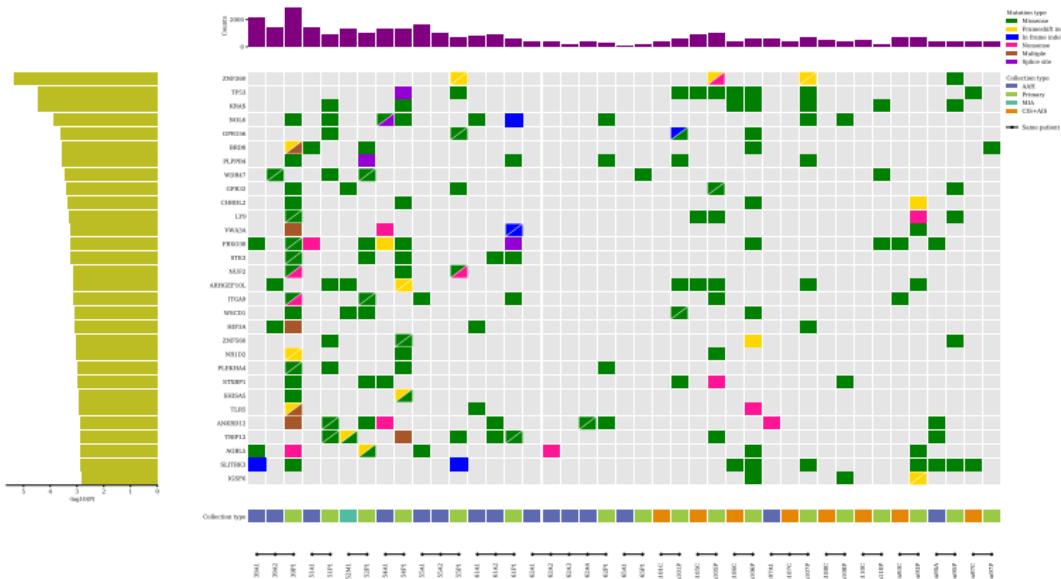


Figure: CoMut Plot with SQC Patients

## Somatic Variant with BWA in ADC



## Figure: CoMut Plot with ADC Patients

# Findings in SNVs Analysis

# Results

## VAF Analysis

# VAF?

# Findings in VAF Analysis

# Results

## Differences in Gene Expression Levels

# RSEM?

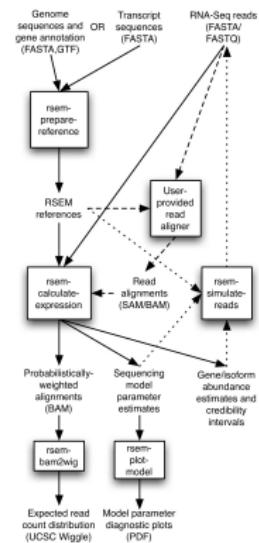


Figure: RSEM workflow (B. Li & Dewey, 2011)

# DESeq2

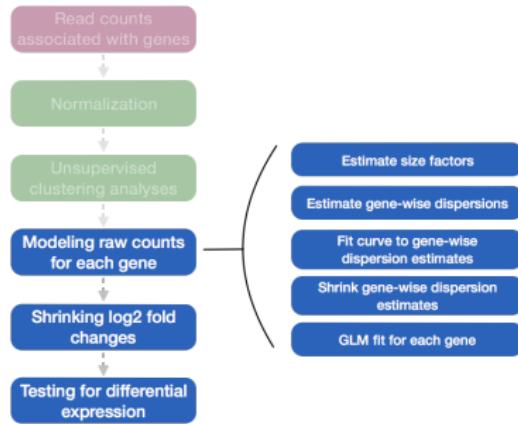


Figure: DESeq2 workflow (Love et al., 2014)

# DEG Selection Strategy

DEG: differentially expressed genes

Fold Change

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

P-value

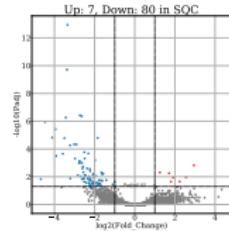
$$P\text{-value} < 0.05$$

Adjusted P-value

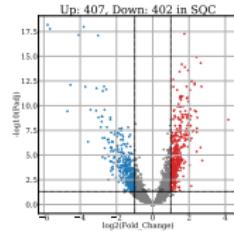
$$P_{adj} < 0.05$$

# DEG Volcano Plots in SQC

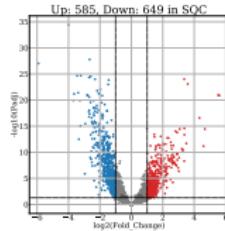
Normal → Dysplasia → CIS → Primary (SQC)



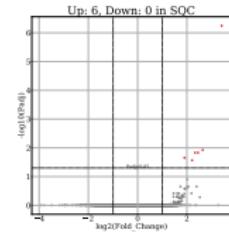
(a) Normal-Dysplasia



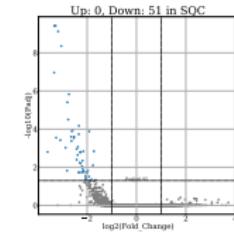
(b) Normal-CIS



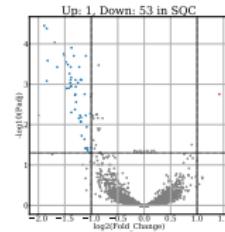
(c) Normal-Primary



(d) Dysplasia-CIS



(e) Dysplasia-Primary

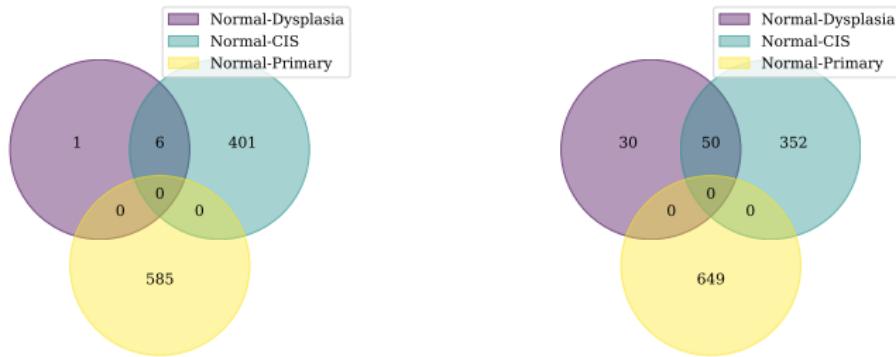


(f) CIS-Primary

Figure: DEG Volcano Plots in SQC

# DEG Venn Diagram with Bowtie2 in SQC

Normal → Dysplasia → CIS → Primary (SQC)



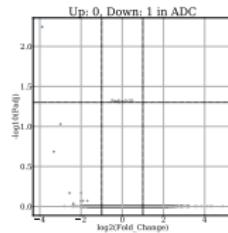
(a) Up-regulated

(b) Down-regulated

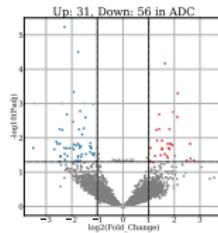
Figure: DEG Venn Diagram in SQC

# DEG Volcano Plots with Bowtie2 in ADC

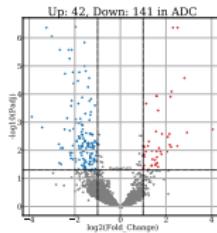
Normal → AAH → AIS → MIA → Primary (ADC)



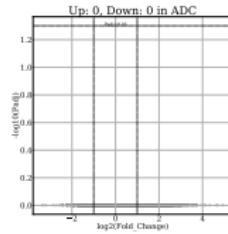
(a) Normal-AAH



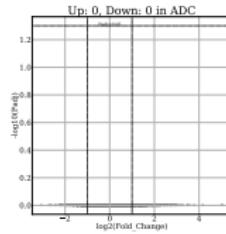
(b) Normal-AIS



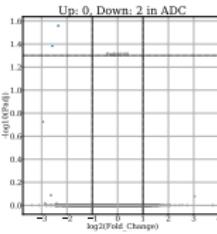
(c) Normal-Primary



(d) AAH-AIS



(e) AAH-Primary

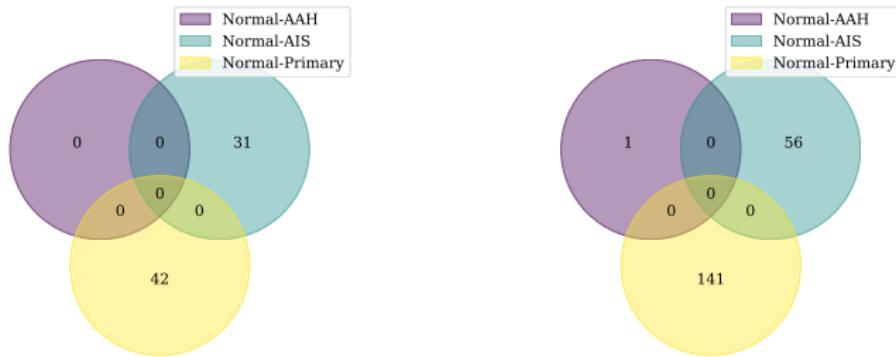


(f) AIS-Primary

Figure: DEG Volcano Plots in ADC

# DEG Venn Diagram with Bowtie2 in ADC

Normal → AAH → AIS → MIA → Primary (ADC)



(a) Up-regulated

(b) Down-regulated

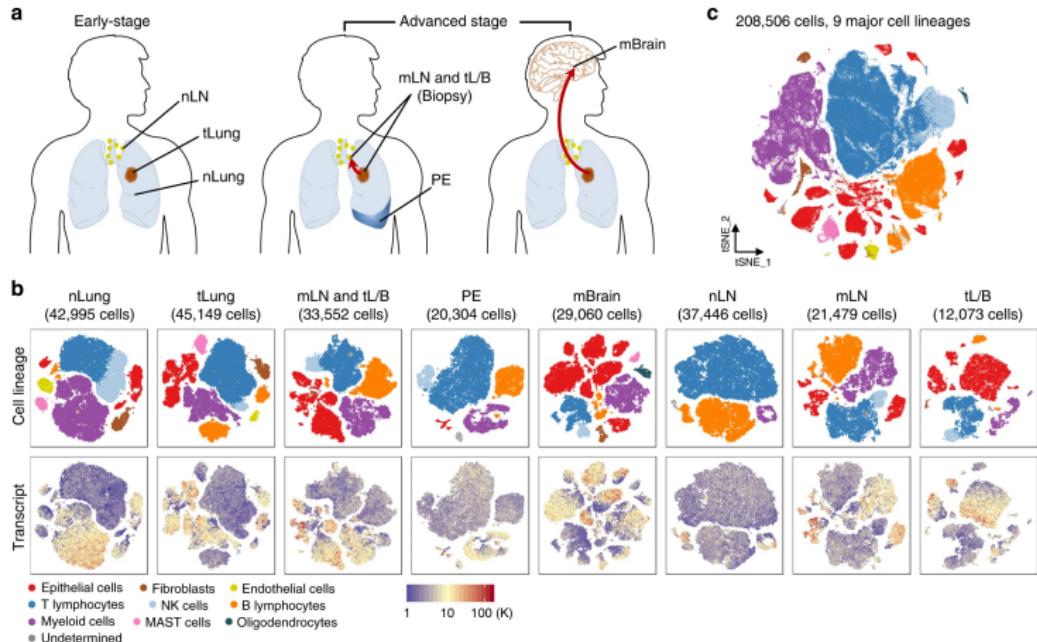
Figure: DEG Venn Diagram in ADC

# Findings in DEG Analysis

# Results

## 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)

# CIBERSORTx

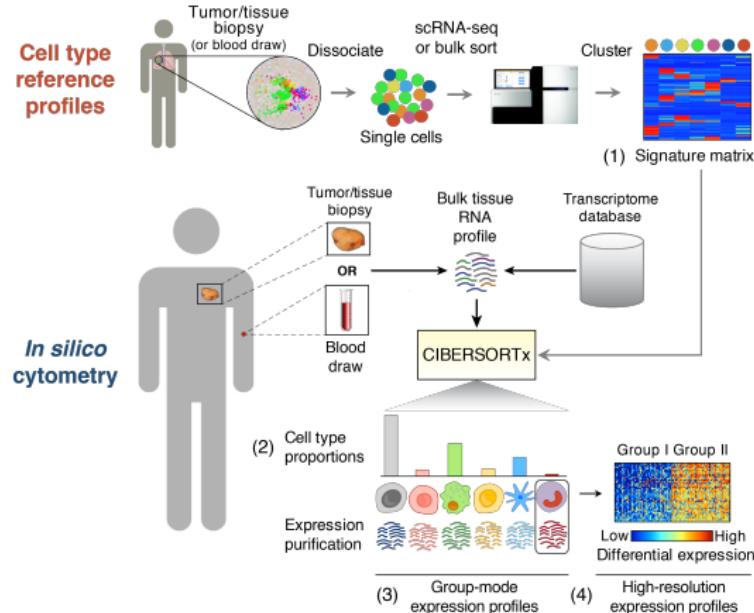


Figure: Workflow for CIBERSORTx (Steen et al., 2020; Newman et al., 2019)

# Cluster Plot with Bowtie2 in ADC

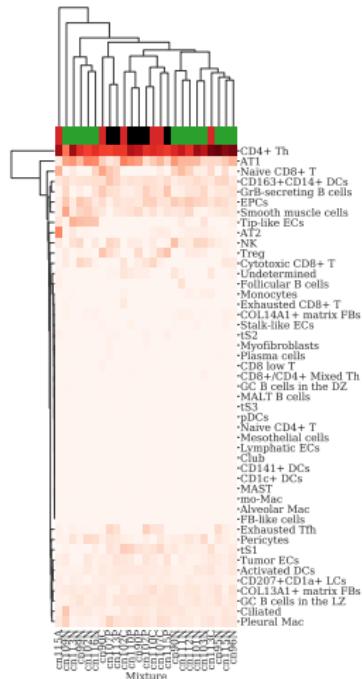


Figure: Cluster Plot in ADC

# Cluster Plot with Bowtie2 in SQC

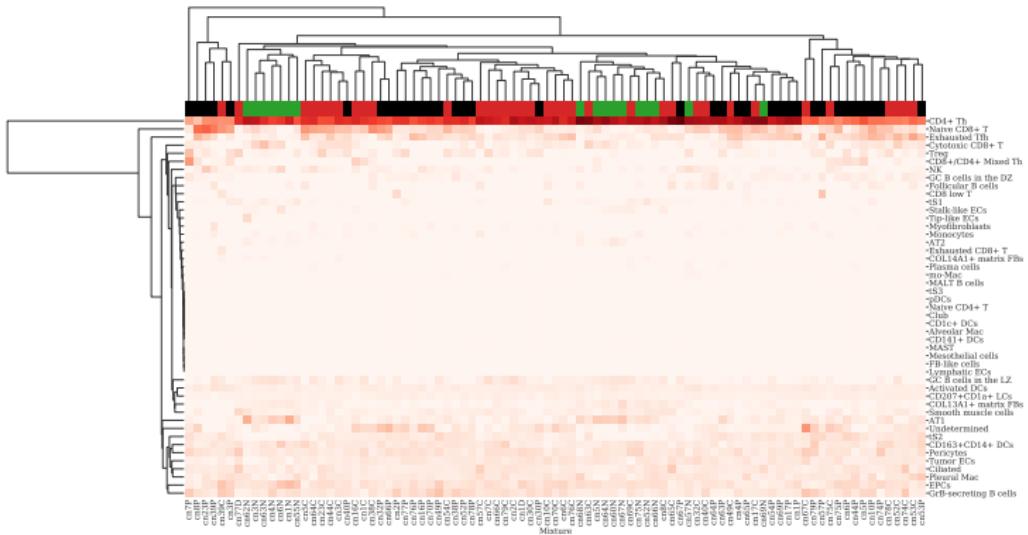


Figure: Cluster Plot in SQC

# Benchmarking of Cell Deconvolution Tools

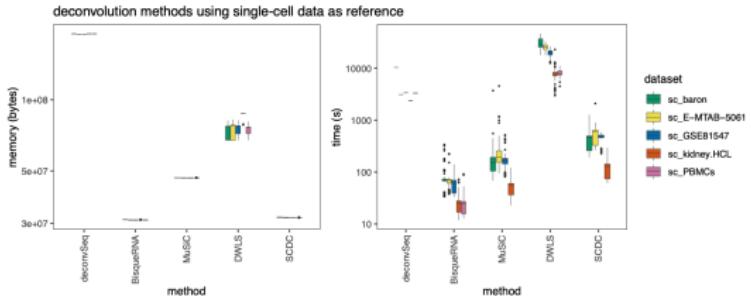
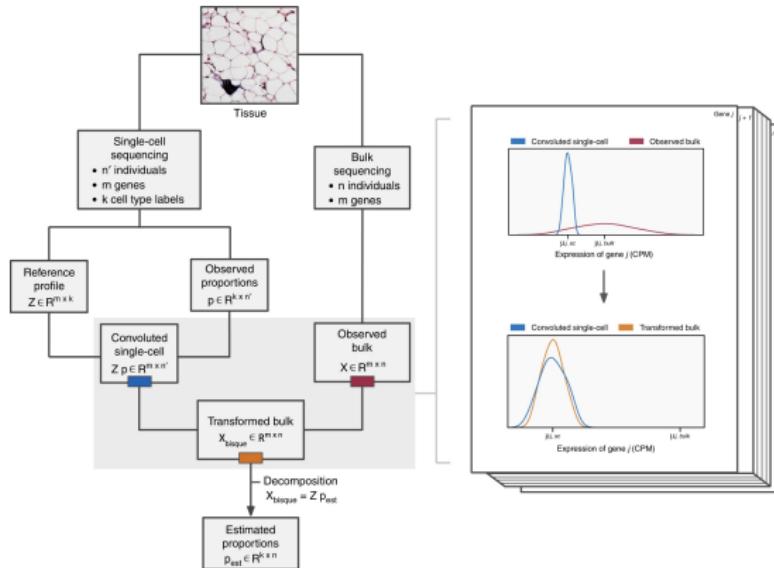


Figure: Memory and time requirements for the cell deconvolution methods (Cobos et al., 2020)

## Top 3 Methods

- ① BisqueRNA (Jew et al., 2020)
- ② MuSiC (X. Wang, Park, Susztak, Zhang, & Li, 2019)
- ③ SCDC (Dong et al., 2021)

# BisqueRNA?



**Figure:** Graphical overview of the Bisque decomposition methods (Jew et al., 2020)

# MuSiC?

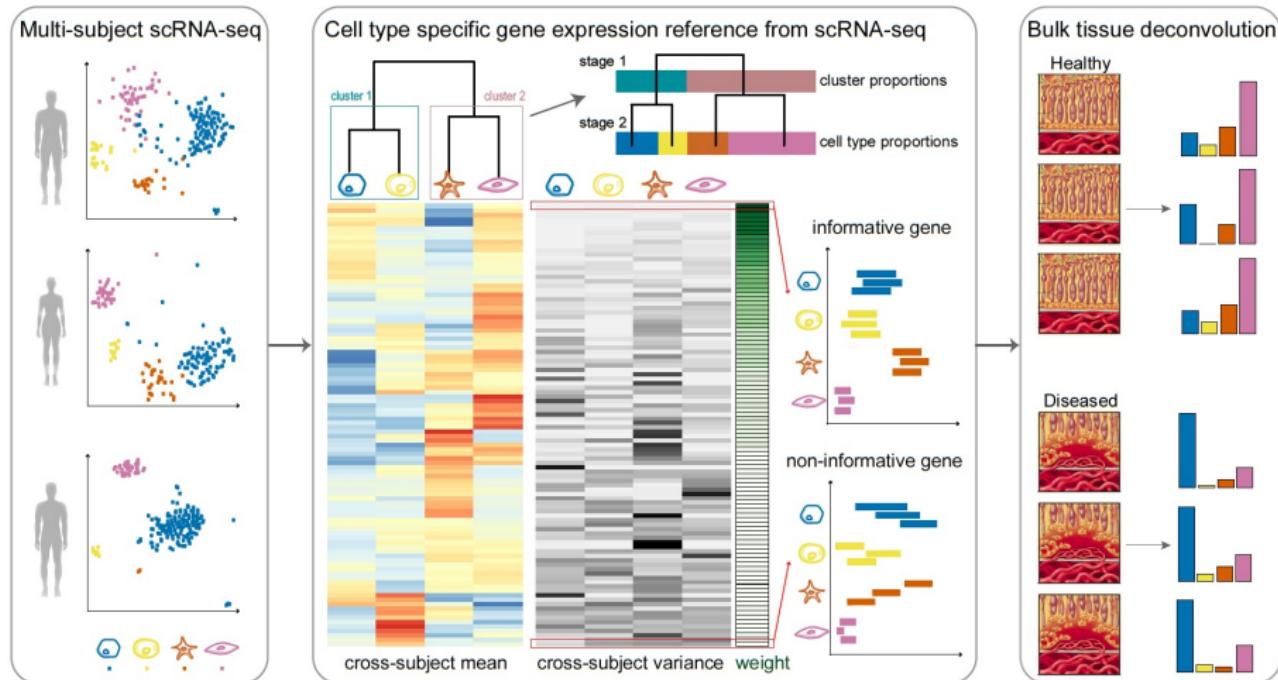
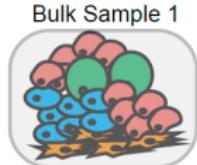


Figure: Overview of MuSiC framework (X. Wang et al., 2019)

# SCDC?

Bulk RNA-seq



Bulk Sample 2

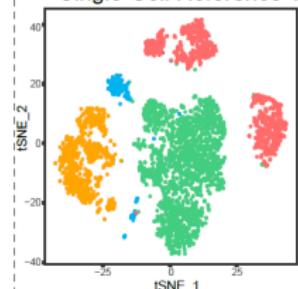


Bulk Sample 3

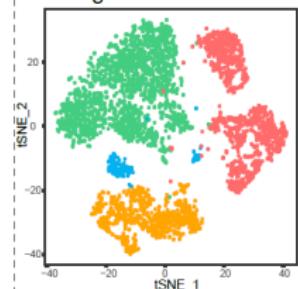


scRNA-seq

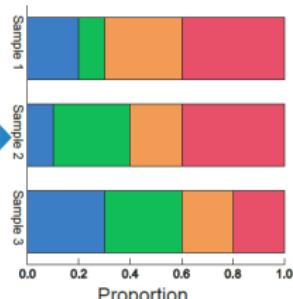
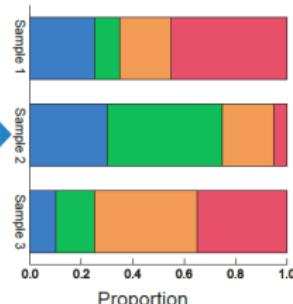
Single-Cell Reference 1



Single-Cell Reference 2



Deconvolution



ENSEMBLE

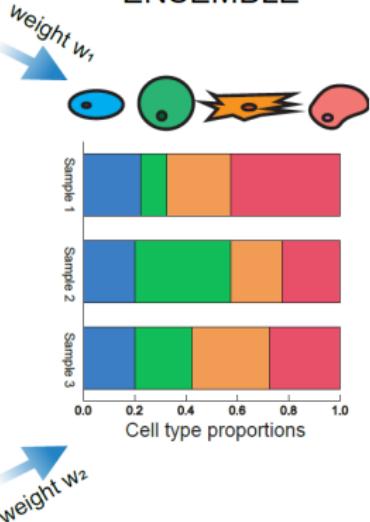


Figure: Overview of deconvolution by SCDC (Dong et al., 2021)

# Findings in Bulk Cell Deconvolution

# Results

## Tumor Evolution Trajectories Analysis

# Revolver?

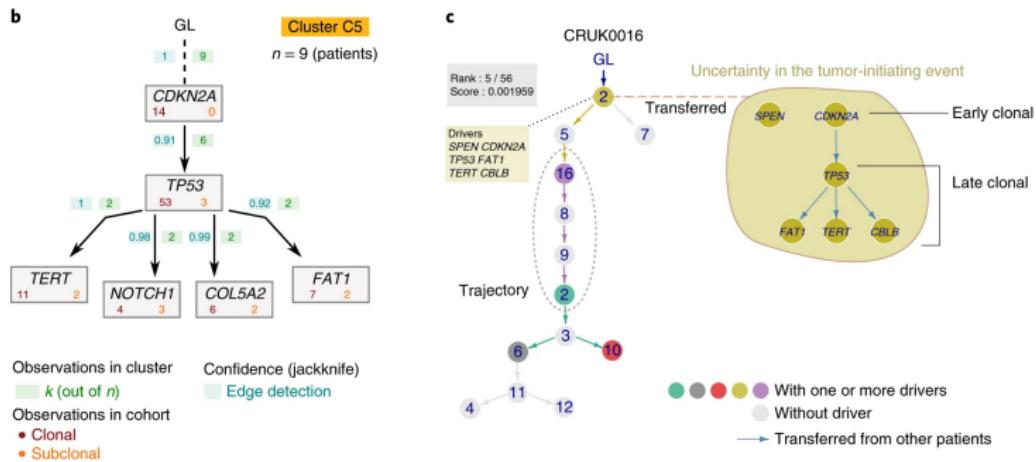


Figure: Repeated Evolutionary Trajectories (Caravagna et al., 2018)

# Findings in Tumor Evolution Trajectories Analysis

# Discussion

## 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.
- Cobos, F. A., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P., & De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11(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

- Crowdis, J., He, M. X., Reardon, B., & Van Allen, E. M. (2020). Comut: visualizing integrated molecular information with comutation plots. *Bioinformatics*, 36(15), 4348–4349.
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... Li, H. (2021, 02). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). Retrieved from <https://doi.org/10.1093/gigascience/giab008> (giab008) doi: 10.1093/gigascience/giab008
- 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.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1), 15–21.

## References III

- Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C. M., Zou, F., & Jiang, Y. (2021). Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, 22(1), 416–427.
- 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.
- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *IEEE Annals of the History of Computing*, 9(03), 90–95.

## References IV

- Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K. M., ... Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1), 1–11.
- Jiang, L., Zheng, J., Kwan, J. S., Dai, S., Li, C., Li, M. J., ... others (2019). Witer: a powerful method for estimation of cancer-driver genes using a weighted iterative regression modelling background mutation counts. *Nucleic acids research*, 47(16), e96–e96.
- Kandoth, C., Gao, J., qwangmsk, Mattioni, M., Struck, A., Boursin, Y., ... Chavan, S. (2018, February). *mskcc/vcf2maf: vcf2maf v1.6.16*. Zenodo. Retrieved from  
<https://doi.org/10.5281/zenodo.1185418> doi:  
10.5281/zenodo.1185418

## References V

- 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.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4), 357.
- 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.
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14), 1754–1760.

## References VI

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–2079.
- 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.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., ... Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, 17(1), 1–14.
- Minna, J. D., Roth, J. A., & Gazdar, A. F. (2002). Focus on lung cancer. *Cancer cell*, 1(1), 49–52.
- Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., ... others (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7), 773–782.

## References VII

- Nilsen, G., Liestol, K., & Lingjaerde, O. (2013). *Copynumber: Segmentation of single-and multi-track copy number data by penalized least squares regression. R package version, 1(0)*.
- Nilsen, G., Liestøl, K., Van Loo, P., Vollan, H. K. M., Eide, M. B., Rueda, O. M., ... others (2012). Copynumber: efficient algorithms for single-and multi-track copy number segmentation. *BMC genomics*, 13(1), 1–16.
- pandas development team, T. (2020, February). *pandas-dev/pandas: Pandas*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.3509134> doi: 10.5281/zenodo.3509134
- Picard toolkit. (2019). <http://broadinstitute.github.io/picard/>. Broad Institute.

## References VIII

- 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.
- Steen, C. B., Liu, C. L., Alizadeh, A. A., & Newman, A. M. (2020). Profiling cell type abundance and expression in bulk tissues with cibersortx. In *Stem cell transcriptional networks* (pp. 135–157). Springer.
- Travis, W. D. (2002). Pathology of lung cancer. *Clinics in chest medicine*, 23(1), 65–81.
- 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 IX

- 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.
- Waskom, M., & the seaborn development team. (2020, September). *mwaskom/seaborn*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.592845> doi: 10.5281/zenodo.592845

# References X

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 56 - 61). doi: 10.25080/Majora-92bf1922-00a