Lung Precancer Analysis

Jaewoong Lee Yeonsong Choi Ilsun Yun Semin Lee

Department of Biomedical Engineering
Ulsan National Institute of Science and Technology

jwlee230@unist.ac.kr

2021-05-18

Overview

Introduction

Lung Cancer

- Squamous cell carcinoma
- Adenocarcinoma

Precancer

Introduction

Study Objectives

Study Objectives

- Find different mutations
  - between WES
  - between WTS
  - from cancer
  - from precancer
- Pathway examine from the mutations
  - of WES
  - of RNA-seq
- Ultra-deep sequencing to find an *infinitesimal* quantity of Non-Circulating Tumor DNA
  - from blood
  - from urine
  - frrom bronchus
- Diagnostic performace

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 - Dysplasia - CIS - SQC (n=80)
  2. Normal - AAH - AIS - MIA - ADC (n=28)

Methods

Methods

Workflows

Data pre-processing for variant discovery

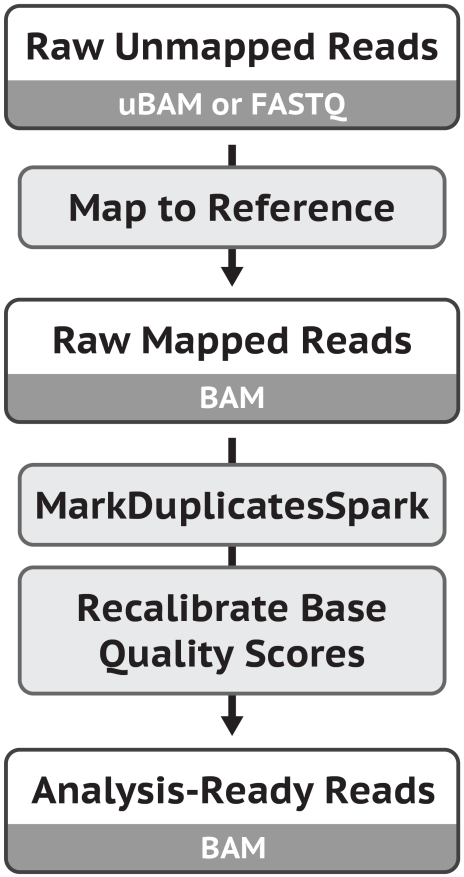


Figure: Data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

Somatic short variant discovery

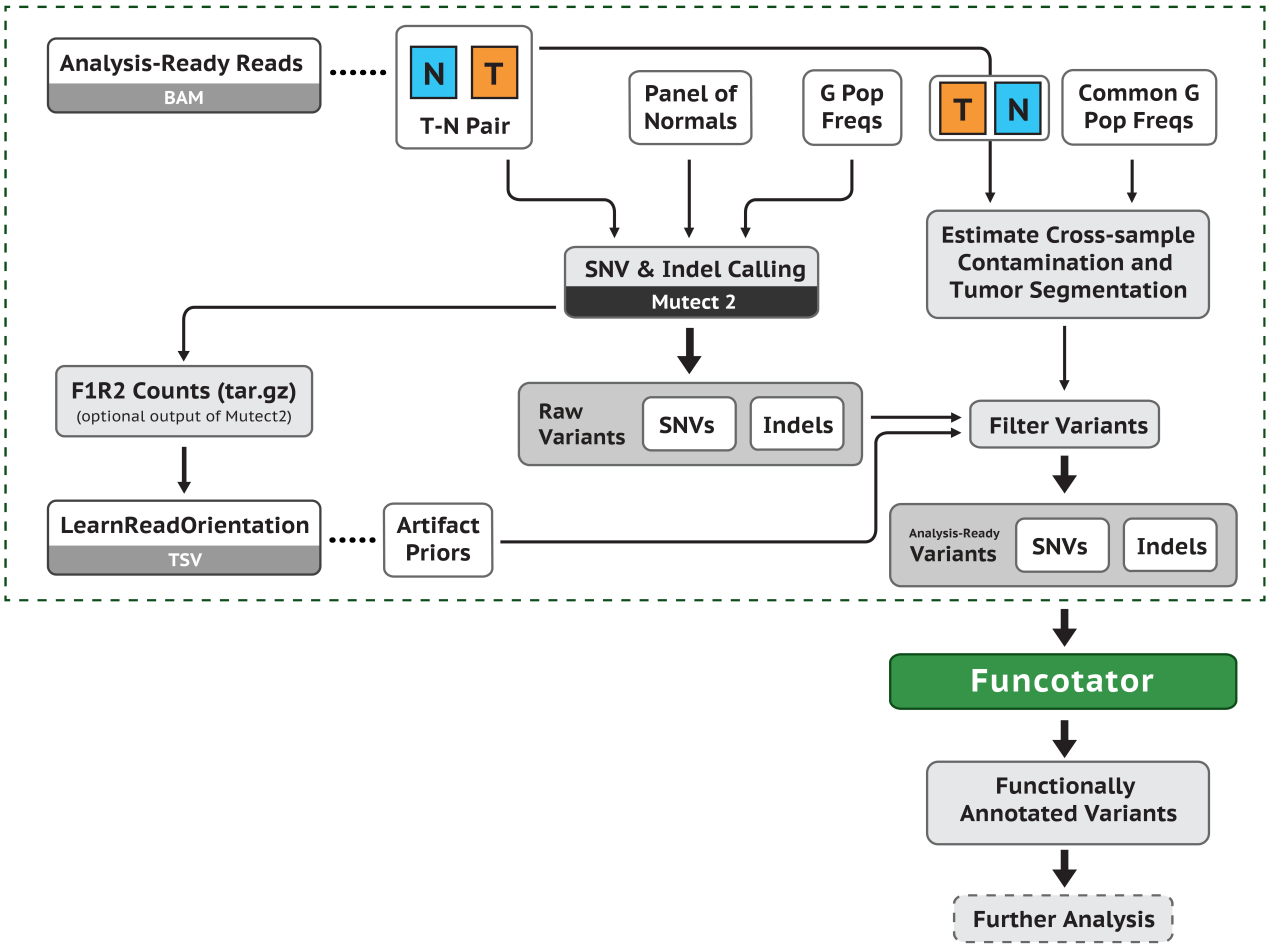


Figure: Somatic short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

Germline short variant discovery

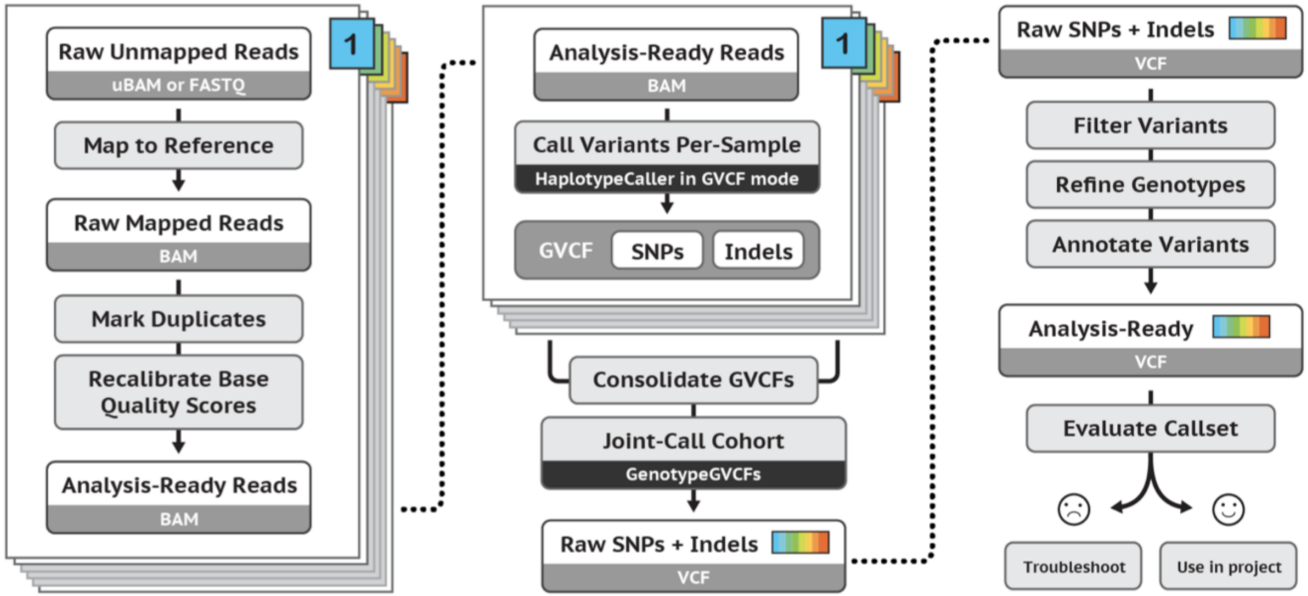


Figure: Germline short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

RNA-seq short variant discovery

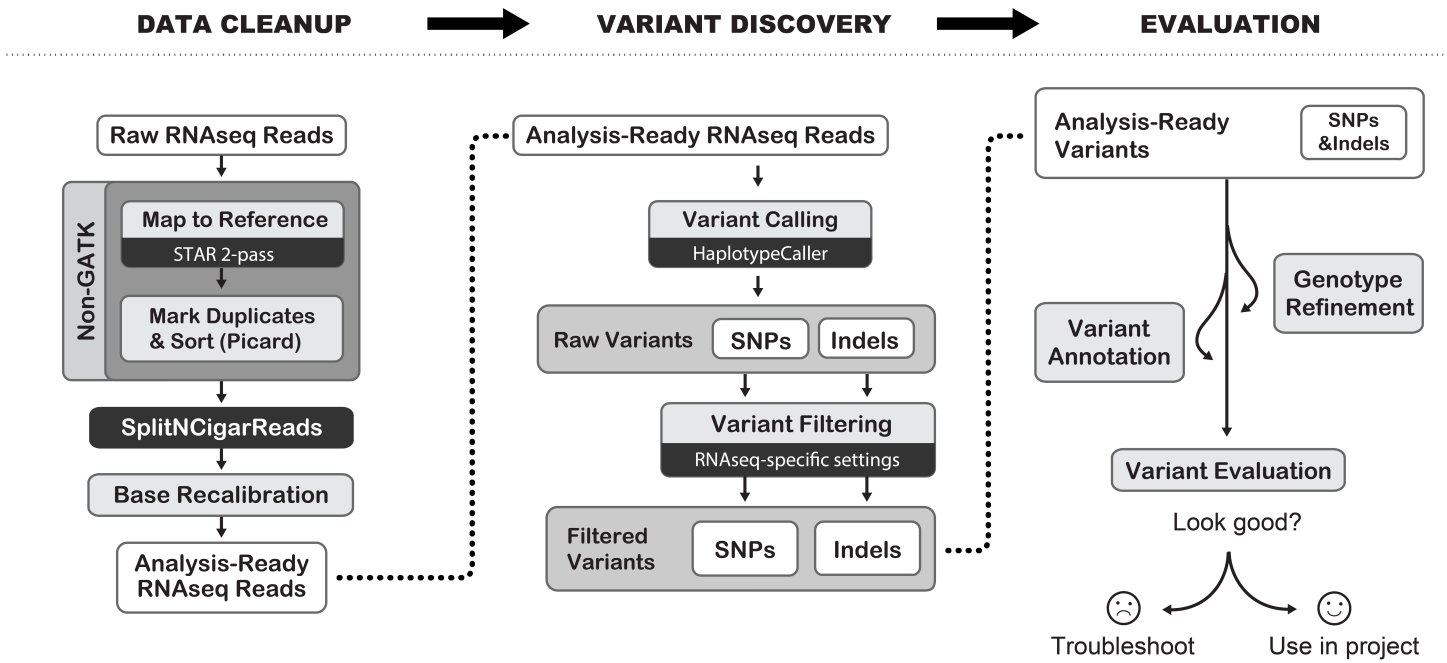


Figure: RNA-seq short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

Methods

Miscellaneous

Used Bioinformatics Tools

- FastQC (Andrews et al., 2012)
- 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)
- RSEM (B. Li & Dewey, 2011)

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)

Results

Results

Quality Checks with FastQC

FastQC?

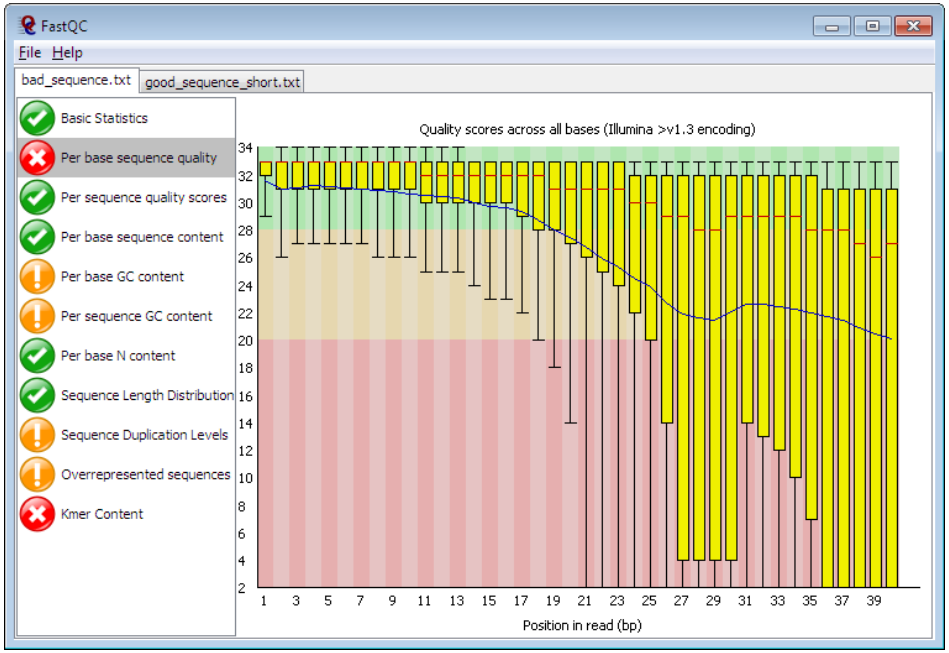


Figure: Example of FastQC Result (Andrews et al., 2012)

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

FastQC on WES

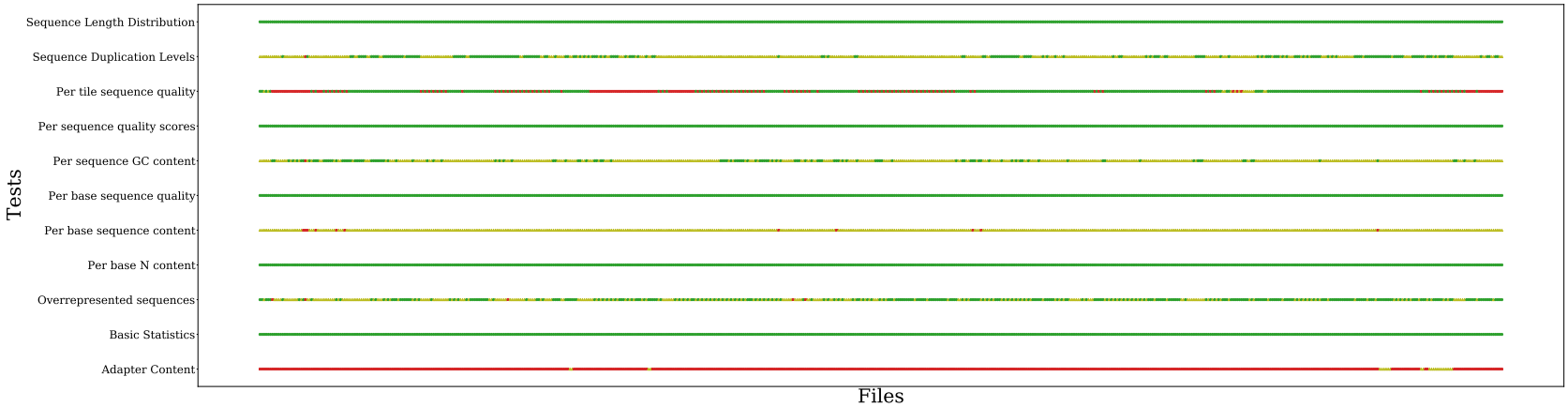
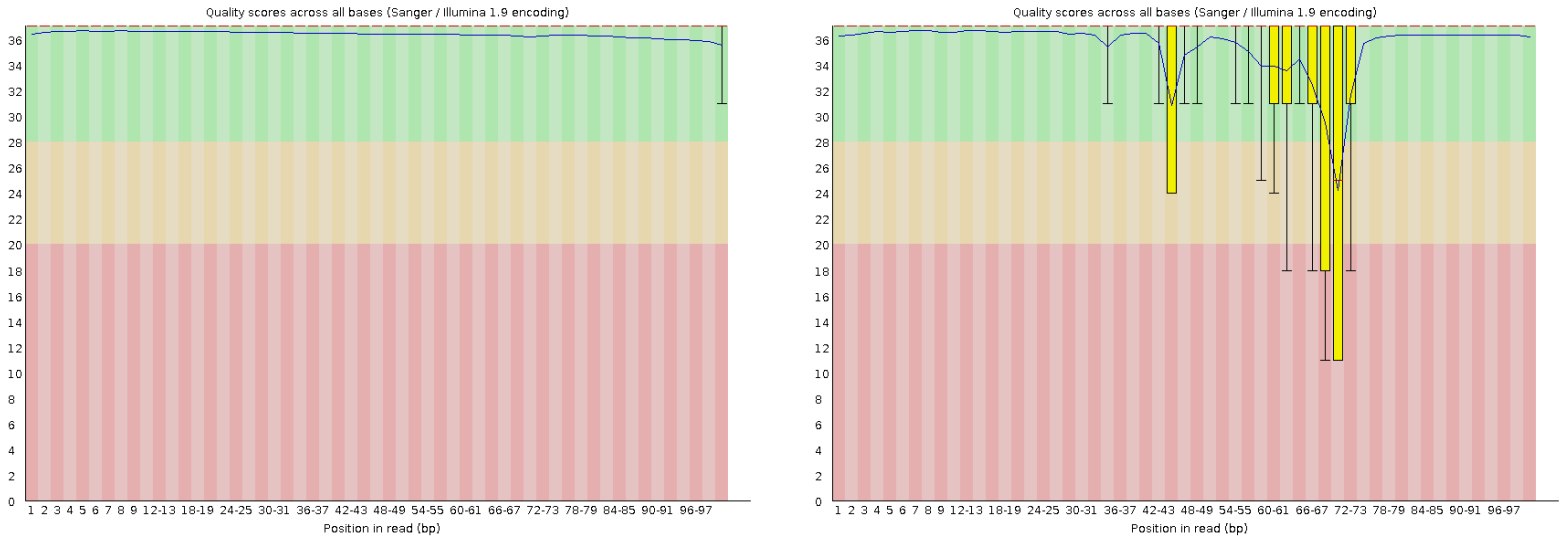


Figure: FastQC with WES data

∴ Only 33P1 has more than 3 failures: 6 FAILs.
∴ 33P1 is excluded at further analysis.

Failure on 33P1 I



(a) 33N (b) 33P1

Figure: Per Base Sequence Quality Results

Failure on 33P1 II

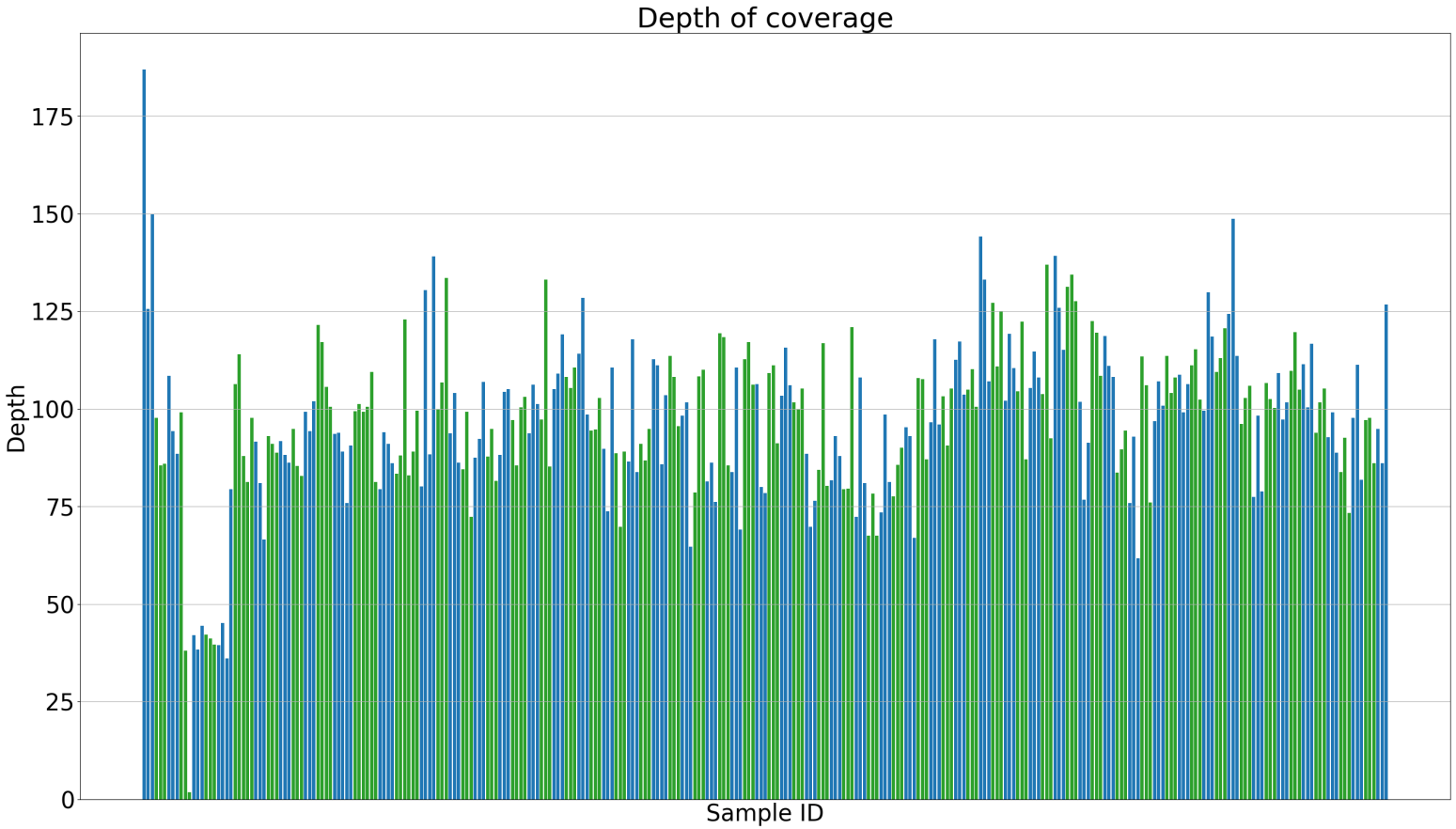


Figure: Coverage Depth Plot

FastQC on WTS

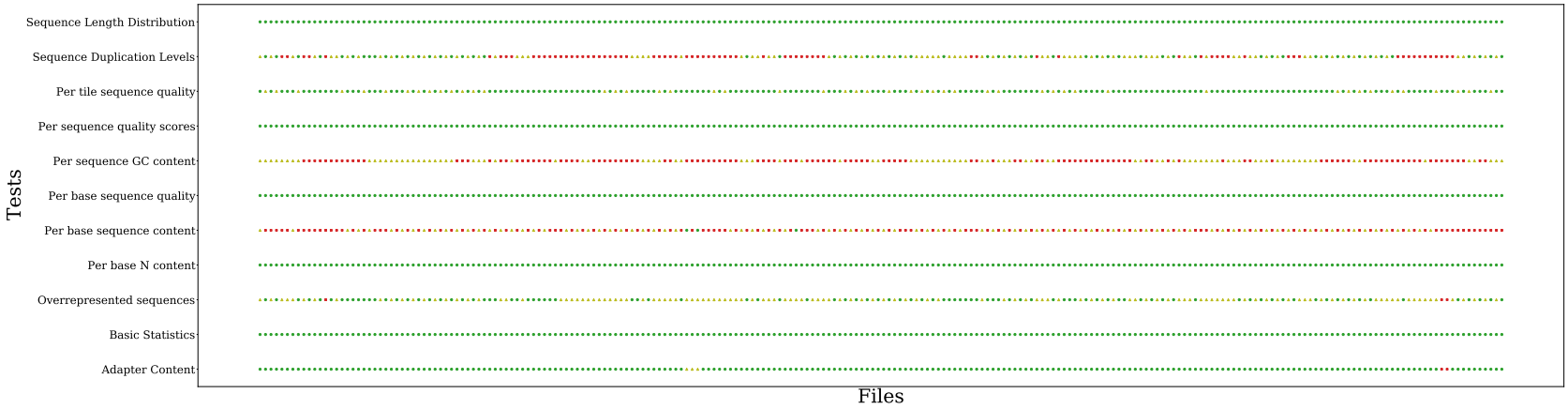


Figure: FastQC with WTS data

$\therefore$ No sample has more than 5 failures.

$\therefore$ All sample are good to analysis.

Results

Quality Checks with Sequenza

Sequenza?

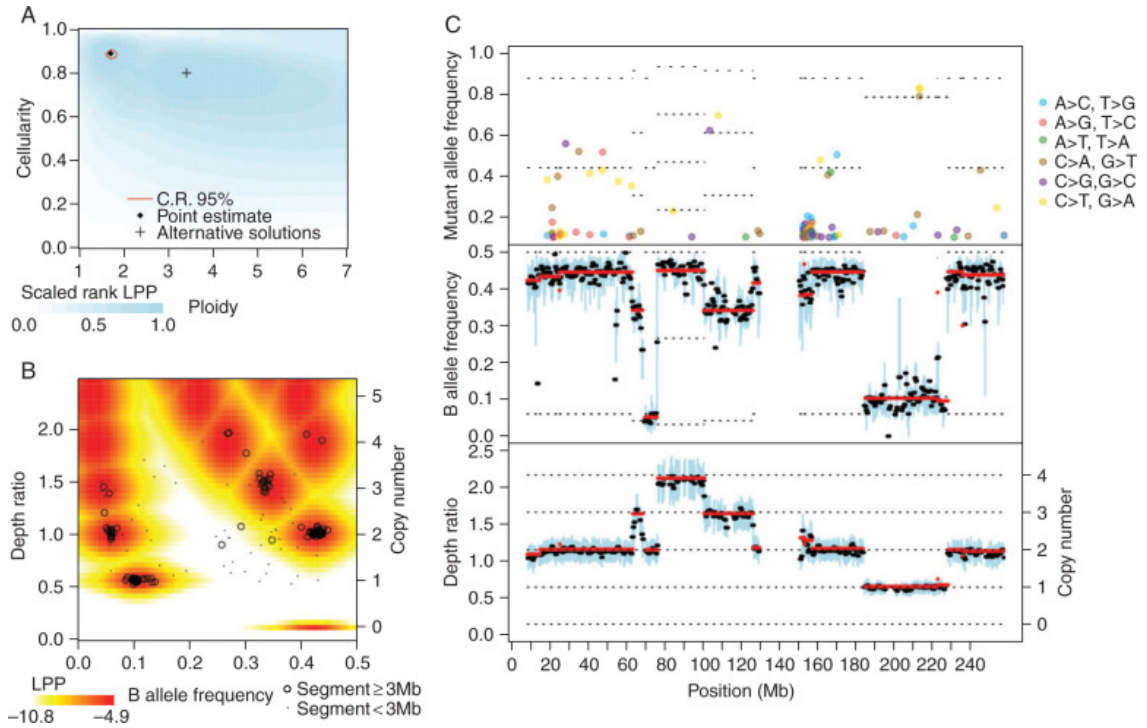
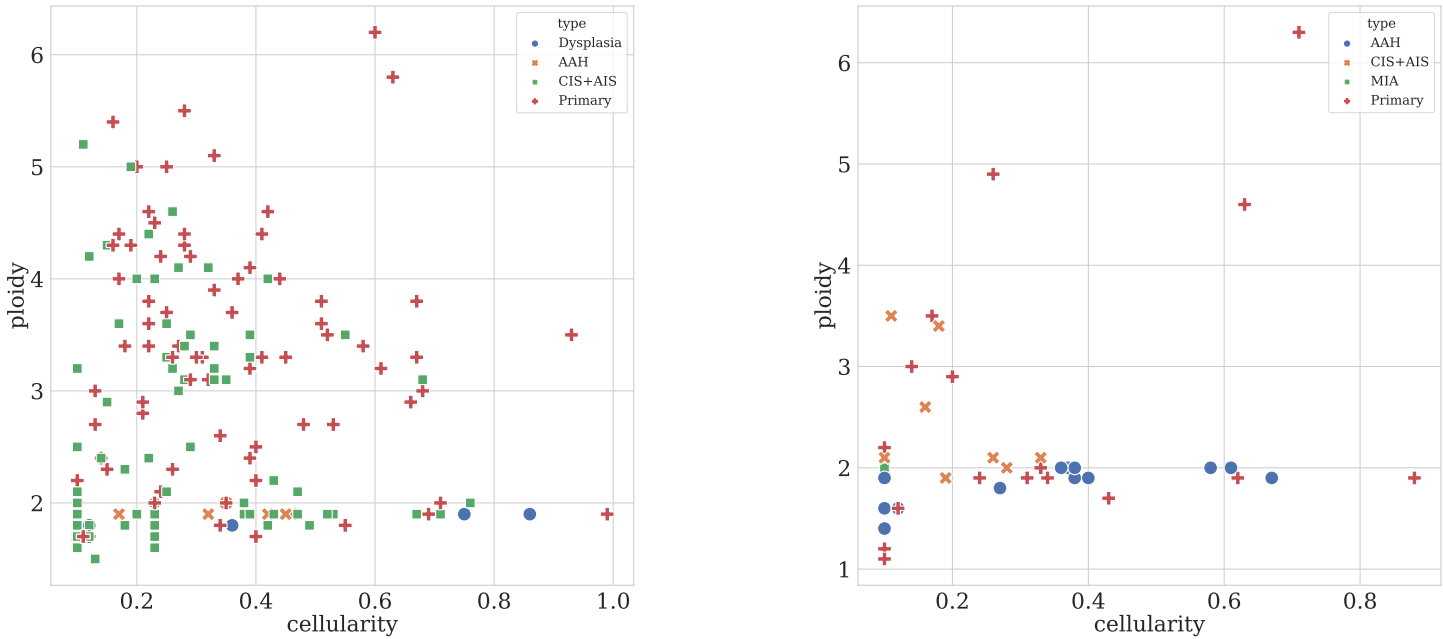


Figure: Representative Output of the Sequenza (Favero et al., 2015)

Cellularity & Ploidy on WES



(a) SQC Samples

(b) ADC Samples

Figure: Cellularity and Ploidy from Sequenza

Genome View on Patient #57

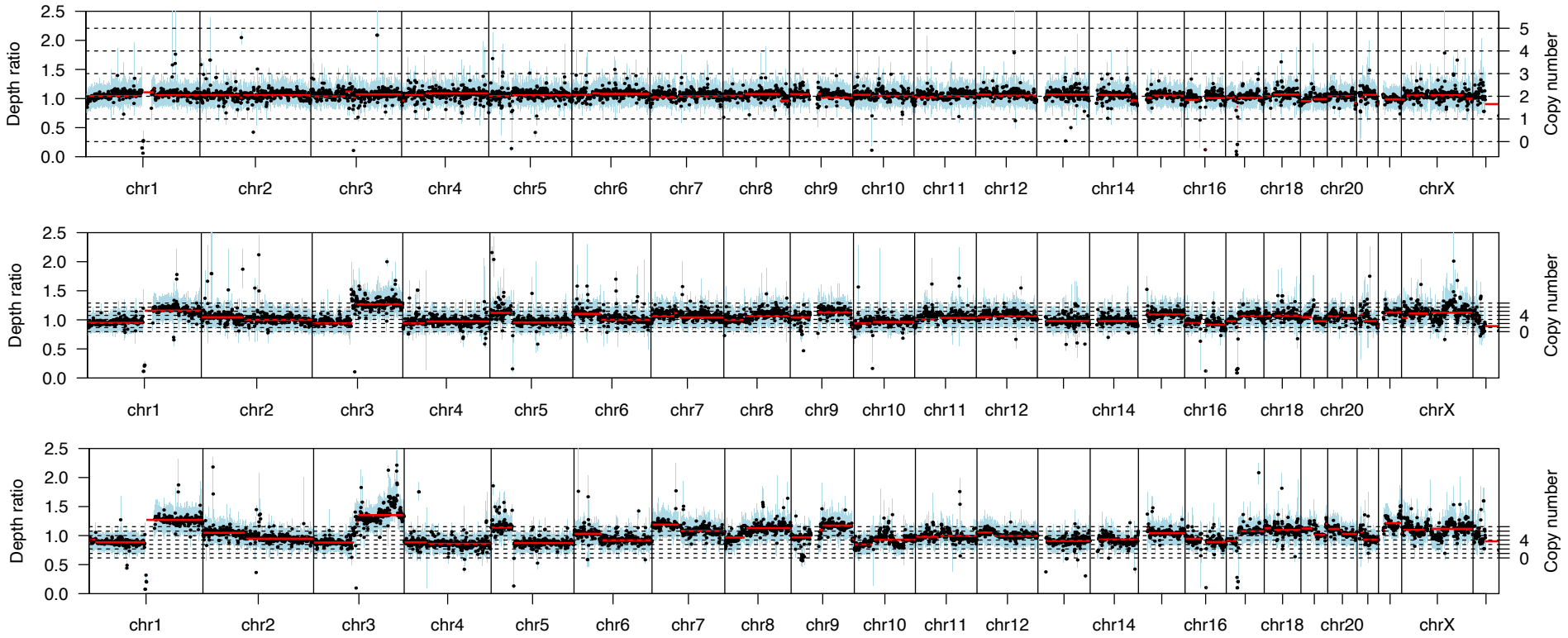


Figure: Dysplasia-CIS-Primary tumor

CNV of SQC

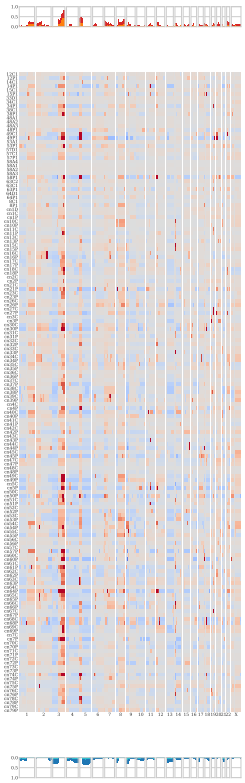


Figure: CNV Plot with SQC Patients

CNV of ADC

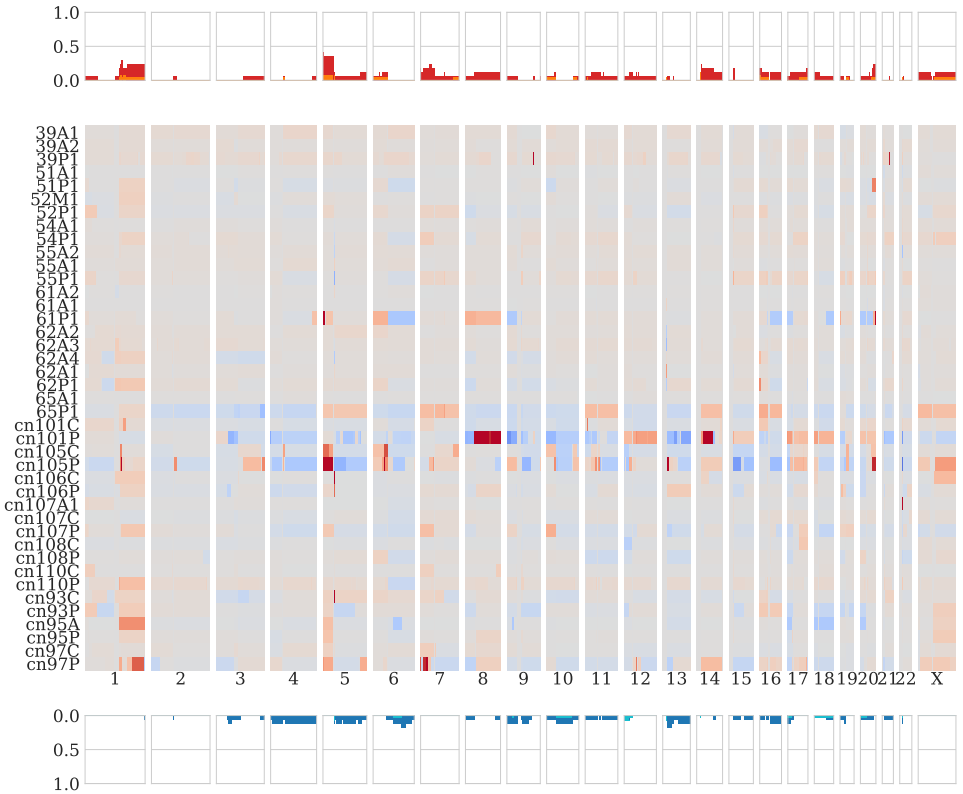


Figure: CNV Plot with ADC Patients

Results

Mutect2

Mutect2?

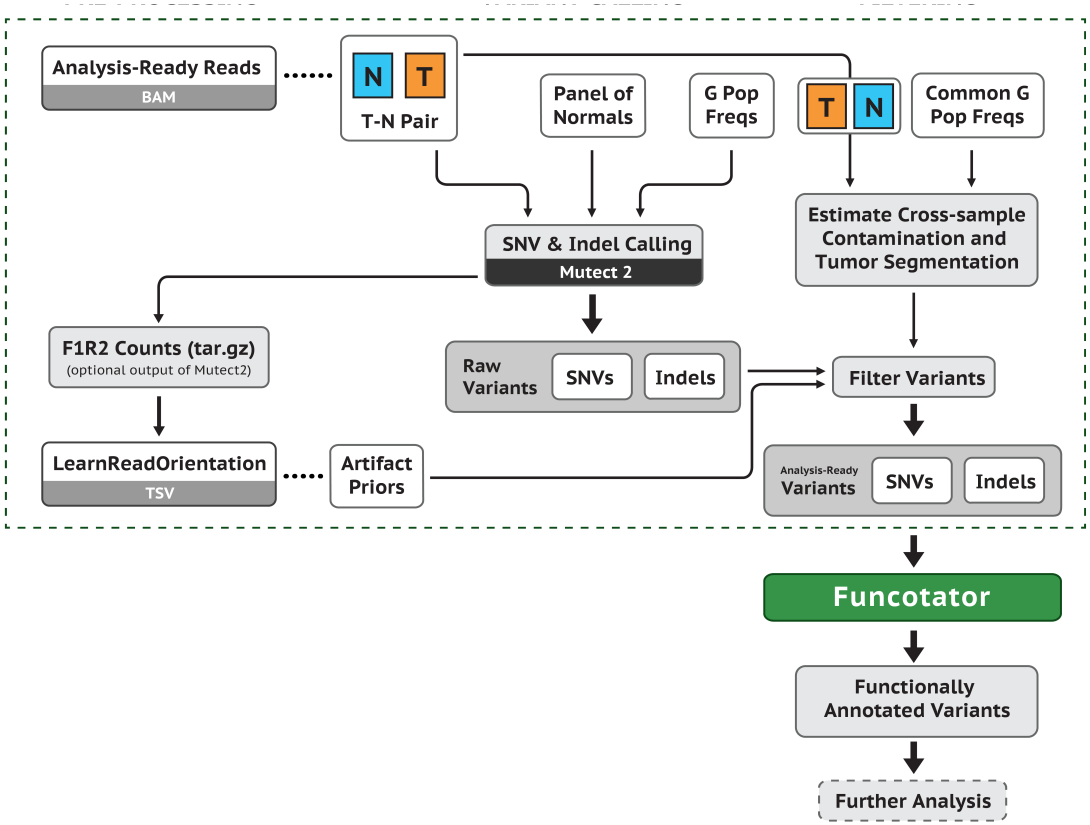


Figure: Somatic short variant discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

Witer?

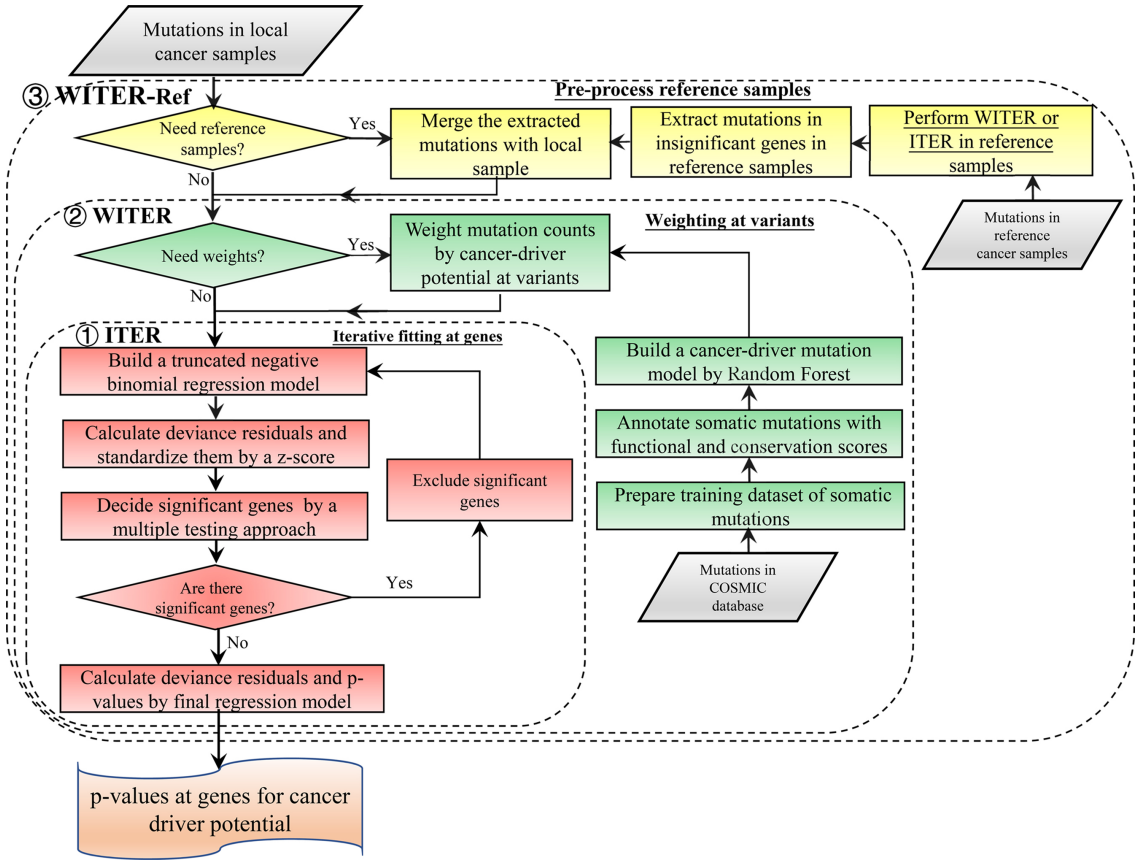


Figure: Witer diagram for detecting cancer-drive genes (Jiang et al., 2019)

Somatic Variant in SQC

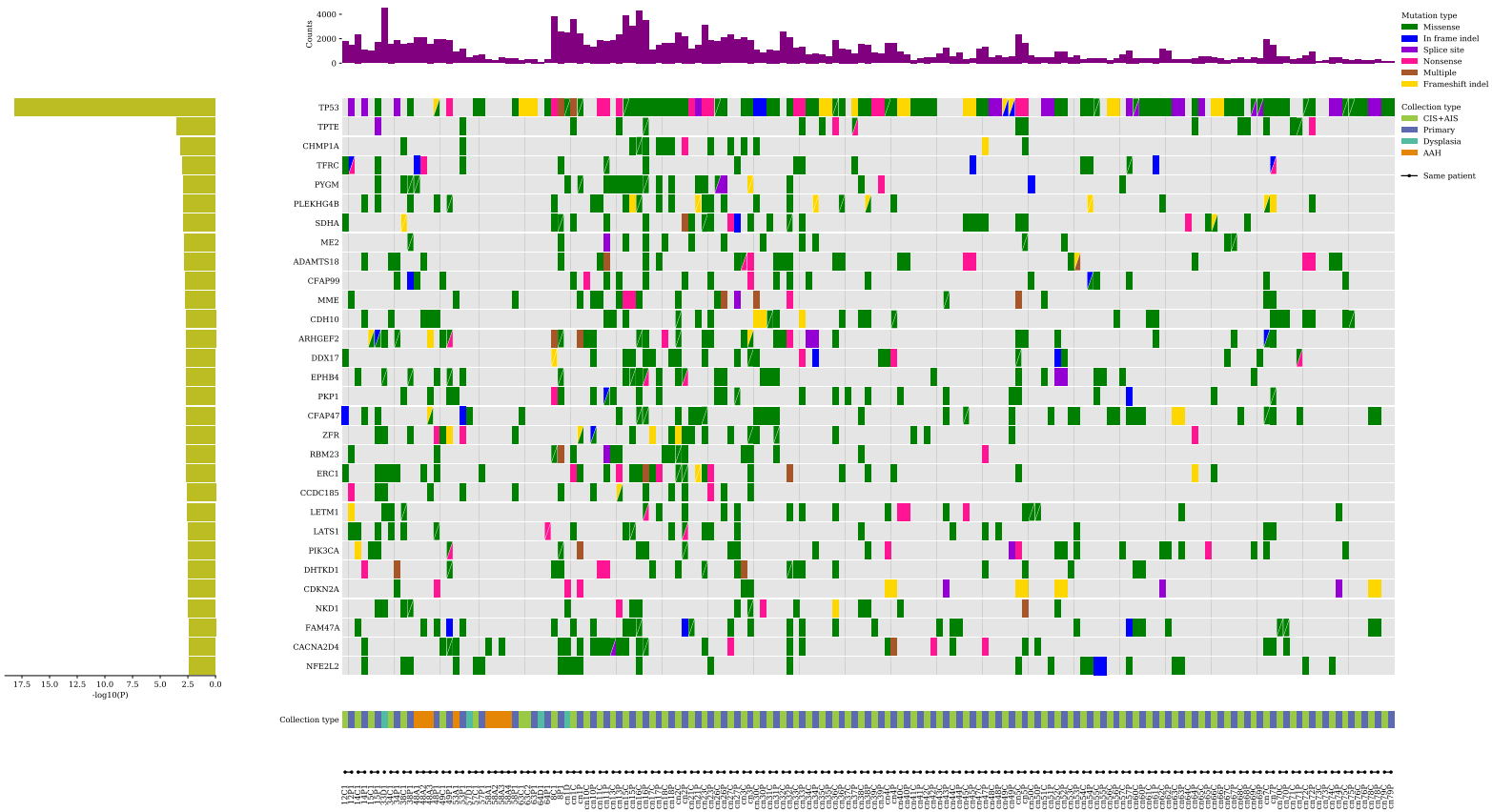


Figure: CoMut Plot with SQC Patients

Somatic Variant in ADC

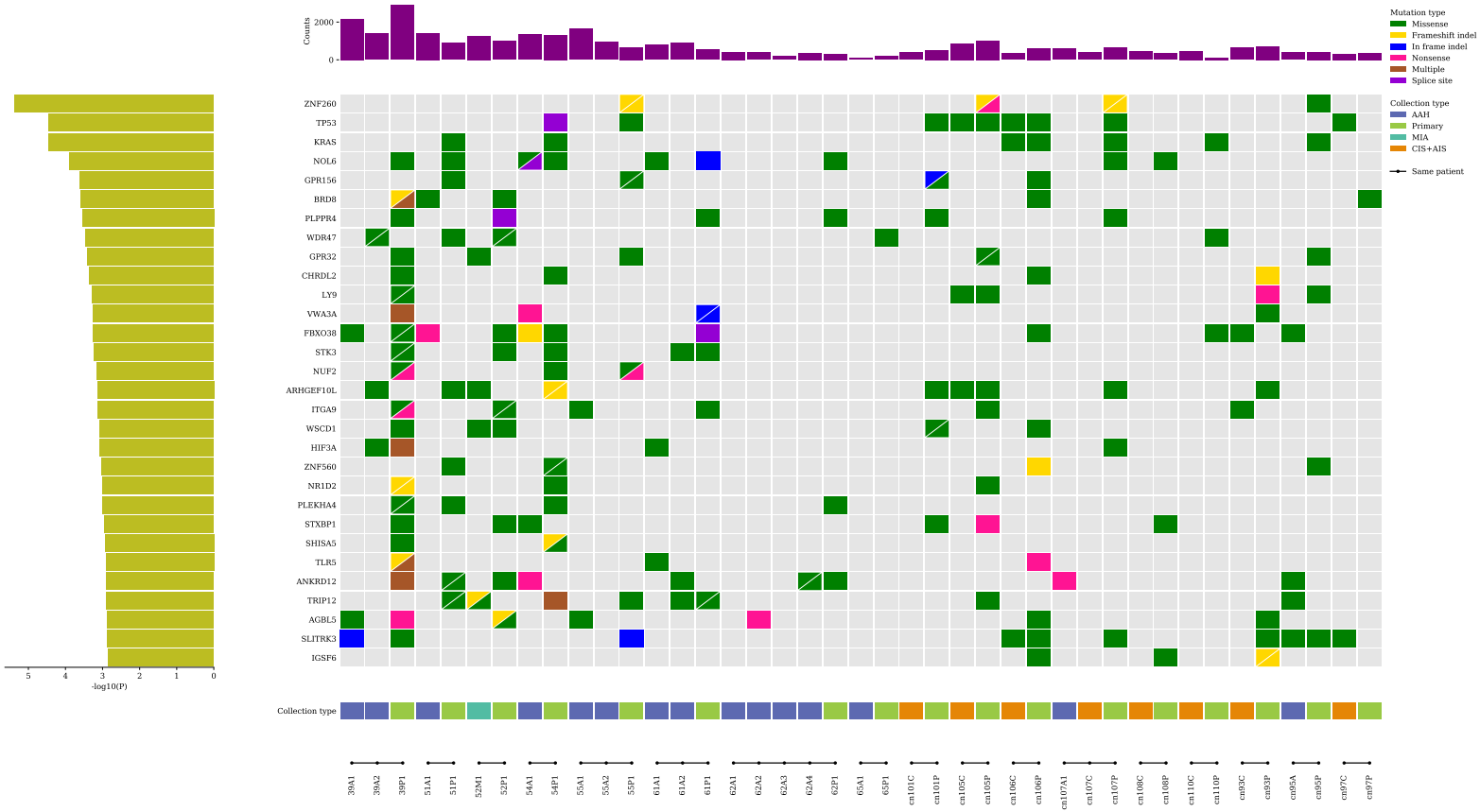


Figure: CoMut Plot with ADC Patients

Results

RSEM

RSEM?

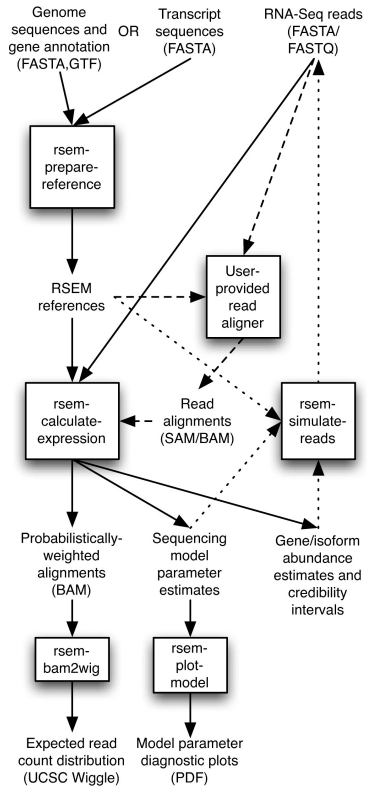


Figure: The RSEM workflow (B. Li & Dewey, 2011)

Volcano Plot in SQC

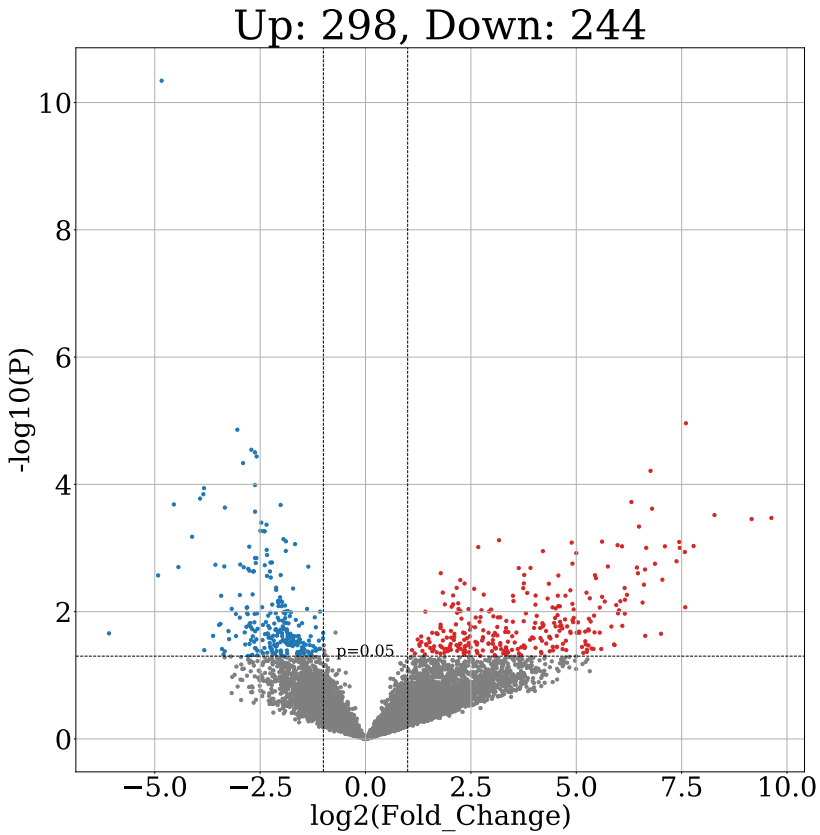


Figure: Volcano Plot in SQC

Volcano Plot in ADC

Discussion

References I

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

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.

References II

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.

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.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *IEEE Annals of the History of Computing*, *9*(03), 90–95.

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.

References III

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

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 IV

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.

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).

References V

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.

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

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

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