Lung Cancer

Jaewoong Lee

2021-02-22

Contents

1 Introduction **3**

2 Materials **3**

3 Methods **3**
3.1 General Usage . . . 3
3.1.1 Genome Analysis Toolkit . . . 3
3.1.2 Samtools . . . 3
3.2 Alignment . . . 3
3.2.1 Burrows-Wheeler Aligner . . . 3
3.2.2 STAR . . . 3
3.3 Quality Check . . . 3
3.3.1 FastQC . . . 3
3.3.2 Sequenza . . . 3
3.4 SNV Detection Workflow . . . 3

4 Results **3**
4.1 FastQC . . . 3

5 Discussion **3**

6 References **3**

List of Tables

List of Figures

1 Data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011) . . . 4
2 Somatic short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011) . . . 4
3 FastQC with WES data . . . 4
4 FastQC with WTS data . . . 4

1 Introduction

2 Materials

3 Methods

3.1 General Usage

3.1.1 Genome Analysis Toolkit

Genome analysis toolkit (GATK) is a software package for variant discovery among sequencing data (Van der Auwera et al., 2013; DePristo et al., 2011).

3.1.2 Samtools

Samtools is a suite software packages for discovering in high-throughput sequencing data (Li et al., 2009).

3.2 Alignment

3.2.1 Burrows-Wheeler Aligner

Burrows-Wheeler Aligner (BWA) is a software package for aligning short-read sequences unto a large reference genome (Li & Durbin, 2009). BWA-MEM is one of the contained algorithms in BWA software package, is a novel algorithm for mapping sequence reads on a large reference genome (Li, 2013).

3.2.2 STAR

STAR is a swift universal RNA-seq alignment tool (Dobin et al., 2013).

3.3 Quality Check

3.3.1 FastQC

FastQC is a software package which aims to provide a productive method to do quality control check on raw sequence data (Andrews et al., 2012).

3.3.2 Sequenza

Sequenza is a software package to investigate genomic sequencing data, such as cellularity and ploidy estimation, from paired normal-tumor samples (Favero et al., 2015).

3.4 SNV Detection Workflow

4 Results

4.1 FastQC

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.

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.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

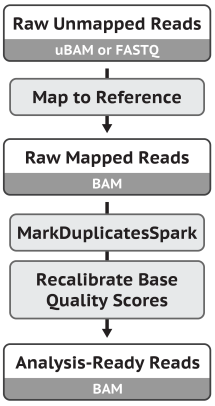

Figure 1: Data pre-processing for variant discovery (Van der Auwera et al., 2013; DePristo et al., 2011)

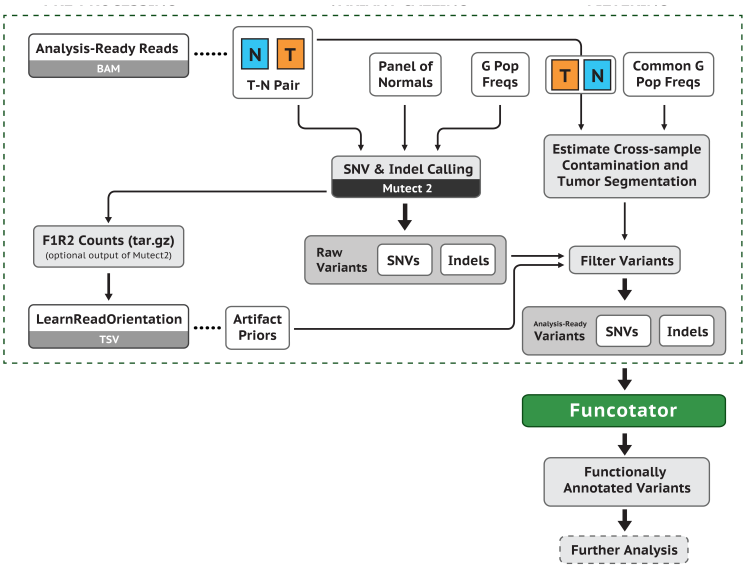

Figure 2: Somatic short variant (SNVs + Indels) discovery workflow (Van der Auwera et al., 2013; DePristo et al., 2011)

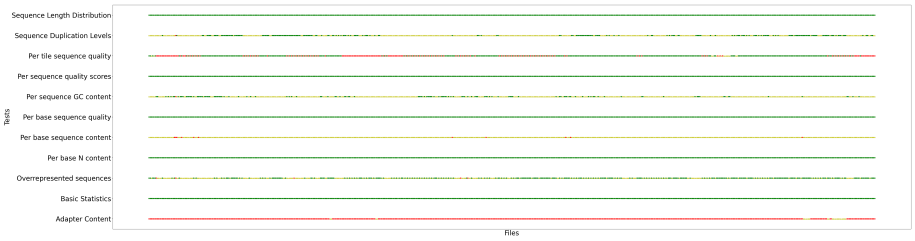

Figure 3: FastQC with WES data

Figure 4: FastQC with WTS data

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, *25*(14), 1754–1760.

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.

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.