## Assessing amplicon files for quality, contamination, and variant identification

## To assess the quality of the files provided, I ran fastqc.

## FASTQC RESULTS

## Basic Statistics

## Table Description automatically generatedTable Description automatically generated

## Per base sequence quality

General quality of the run falling can be observed in both files towards the ends of the run. But read1 looks > read2. To be careful, I did quality trimming to eliminate adapter read-through

Chart

Description automatically generatedChart

Description automatically generated

## Per base sequence content

Quality of the lines (bases) don’t look parallel. Could be due to varying ATGCs (diff in %)

Chart, histogram

Description automatically generatedChart, line chart, histogram

Description automatically generated

## Per sequence GC content



GCs are not normally distributed in aln1 compared to aln2.

Chart, line chart

Description automatically generatedChart, line chart, histogram

Description automatically generated

## Sequence Duplication Levels

% remaining is very low in both reads. This could be due to technical duplicates arising from PCR artefacts or biological duplicates .

Chart

Description automatically generatedChart, line chart

Description automatically generated

## Overrepresented sequences

## Text, application Description automatically generated Text Description automatically generated with medium confidence

## TRIMMING RESULTS

I noticed that the overrepresented seq had primers F& R. To eliminate those read-throughs, I trimmed them.

Number of cores used for trimming: 1

Quality Phred score cutoff: 20

Quality encoding type selected: ASCII+33

Using Illumina adapter for trimming (count: 102). Second best hit was smallRNA (count: 0)

Adapter sequence: 'AGATCGGAAGAGC' (Illumina TruSeq, Sanger iPCR; auto-detected)

Maximum trimming error rate: 0.1 (default)

Minimum required adapter overlap (stringency): 1 bp

Minimum required sequence length for both reads before a sequence pair gets removed: 20 bp

=== Summary ===

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | | aln1.fastq | aln2.fastq |
| Total reads processed: | | | | 575,002 | | 575,002 |
| Reads with adapters: | | | | 110,623 (19.2%) | | 125,683 (21.9%) |
| Reads written (passing filters): | | | | 575,002 (100.0%) | | 575,002 (100.0%) |
| Total basepairs processed: | | | | 86,825,302 bp | | 86,825,302 bp |
| Quality-trimmed: | | | | 732,447 bp (0.8%) | | 1,389,788 bp (1.6%) |
| Total written (filtered): | | | | 85,972,024 bp (99.0%) | | 85,277,341 bp (98.2%) |

aln1.fastq

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 110623 times

Bases preceding removed adapters:

A: 21.4%

**C: 72.2%**

G: 2.4%

T: 4.1%

none/other: 0.0%

aln2.fastq

Sequence: AGATCGGAAGAGC; Type: regular 3'; Length: 13; Trimmed: 125683 times

Bases preceding removed adapters:

A: 16.8%

C: 57.2%

G: 22.4%

T: 3.6%

none/other: 0.0%

Total reads: samtools idxstats aln.sort.bam | awk '{s+=$3+$4} END {print s}' :1150004

Mapped Reads: samtools idxstats aln.sort.bam | awk '{s+=$3} END {print s}' :1052983

UNmapped: 97021 (8.4%)

## READ CONSTITUTION

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1150004 | 0 | total (QC-passed reads + QC-failed reads) | | | |
| 0 | 0 | secondary |  |  |  |
| 0 | 0 | supplementary | |  |  |
| 0 | 0 | duplicates |  |  |  |
| 1052983 | 0 | mapped |  |  |  |
| 91.56% | N/A | mapped % |  |  |  |
| 1150004 | 0 | paired in sequencing | |  |  |
| 575002 | 0 | read1 |  |  |  |
| 575002 | 0 | read2 |  |  |  |
| 1046208 | 0 | properly paired | |  |  |
| 90.97% | N/A | properly paired % | |  |  |
| 1049400 | 0 | with itself and mate mapped | | |  |
| 3583 | 0 | singletons |  |  |  |
| 0.31% | N/A | singletons % | |  |  |
| 646 | 0 | with mate mapped to a different chr | | |  |
| 583 | 0 | with mate mapped to a different chr (mapQ>=5) | | | |

**Maps to EGFR gene (hg19)**

|  |  |  |
| --- | --- | --- |
| chr7 | 55,241,581 | 55,241,794 |
| Chr7 | 55,242,364 | 55,242,579 |
| chr7 | 55,248,948 | 55,249,180 |
| chr7 | 55,259,371 | 55,259,589 |

Read depth of amplicons I calculated manually for amplicons (from samtools depth -I input bam)

**Coverage calculated by samtools -r input.bam**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #rname | startpos | endpos | numreads | covbases | coverage | meandepth | meanbaseq | meanmapq | Input |
| 7 | 55086725 | 55275031 | 477 | 912 | 0.484316 | 0.350932 | 36.2 | 40.2 | Chr7 |
| 7 | 55241581 | 55241794 | 58 | 212 | 99.0654 | 29.9206 | 34.2 | 40 | Amp1 |
| 7 | 55242364 | 55242579 | 167 | 209 | 96.7593 | 107.718 | 36.6 | 39.2 | Amp2 |
| 7 | 55248948 | 55249180 | 151 | 233 | 100 | 92.279 | 36.4 | 40.5 | Amp3 |
| 7 | 55259371 | 55259589 | 101 | 216 | 98.6301 | 65.3699 | 36.1 | 41.3 | Amp4 |

**Variants found**

V1

V2

V1

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Chr** | **Ref** | | **Alt** | | **Func.refGene** | | **Gene.refGene** | **ExonicFunc.refGene** |
| **7** | GAATTAAGAGAAGCA | | - | | exonic | | EGFR | nonframeshift deletion |
| **7** | G | | A | | exonic | | EGFR | synonymous SNV |
| **AAChange** | | | | | | | | |
| EGFR:NM\_001346941:exon13:c.1435\_1449del:p.E479\_A483del,  EGFR:NM\_001346897:exon18:c.2101\_2115del:p.E701\_A705del,  EGFR:NM\_001346899:exon18:c.2101\_2115del:p.E701\_A705del,  EGFR:NM\_001346898:exon19:c.2236\_2250del:p.E746\_A750del,  EGFR:NM\_001346900:exon19:c.2077\_2091del:p.E693\_A697del,  EGFR:NM\_005228:exon19:c.2236\_2250del:p.E746\_A750del | | | | | | | | |
| EGFR:NM\_001346941:exon14:c.G1560A:p.Q520Q,E  GFR:NM\_001346897:exon19:c.G2226A:p.Q742Q,  EGFR:NM\_001346899:exon19:c.G2226A:p.Q742Q,  EGFR:NM\_001346898:exon20:c.G2361A:p.Q787Q,  EGFR:NM\_001346900:exon20:c.G2202A:p.Q734Q,  EGFR:NM\_005228:exon20:c.G2361A:p.Q787Q | | | | | | | | |
| **1000G\_ALL** | **dbSNP** | **COSMIC\_ID** | | **ClinVar\_SIG** | | **ClinVar\_DIS** | | |
| . | rs727504233 | COSM6225 | | drug response | | Tyrosine\_kinase\_inhibitor\_response | | |
| 0.43 | rs1050171 | COSM1451600 | | Benign|Likely benign | | not\_specified|Lung\_cancer | | |

V1

V2

V2

V1

V2

|  |  |  |
| --- | --- | --- |
| **COSMIC\_DIS** | | |
| 1(salivary\_gland);2(breast);469(lung);1(thyroid);1(upper\_aerodigestive\_tract);1(large\_intestine) | | |
| 2(breast);1(haematopoietic\_and\_lymphoid\_tissue);7(lung) | | |
| **ClinVar\_ID** | **ClinVar\_DB** | **Otherinfo** |
| RCV000154199.1 | MedGen | het |
| RCV000038427.3|RCV000321080.1 | MedGen|MedGen:OMIM:SNOMED\_CT | het |

**IGV Viewer – Bam input**

A picture containing chart

Description automatically generated

V1

Timeline

Description automatically generated with medium confidence

V2

Some more conclusions:

1. Coverage varies among amplicons in this sample set could be attributed to read depth, GC% changes,
2. The unmapped reads are possible contamination.
3. Several ways to evaluate which variants is more real than others:
4. MAF (with an appropriate cut off threshold)
5. IGV viewer gives the % and map quality for the variants; we can interpret with the ref base.
6. Ref/Alt read ratio (AD) in the vcf file gives an idea of the authenticity
7. Filter by clinical relevance, conservation, SIFT/Polyphen, region (exonic/intronic), function of the gene relevant to disease, genotype (ex: mendelian models), cDNA/AA change (type of change – missense, frameshift etc), publications.