CGAP quality controls

Release v2

Aug 21, 2020

Contents

This is a documentation for quality controls that are part of CGAP pipelines.

CHAPTER 1

Contents

1.1 Docker Image

- The current docker image is `cgap/qc:v2`

The image contains (but is not limited to) the following software packages:

- granite (0.1.0)
- samtools (1.9)
- gatk4 (4.1.2.0)
- picard (2.20.2)
- bamqc.py
- pigz (2.4)
- pbgzip (2b09f97)
- parallel

1.2 VCF Quality Control

1.2.1 Overview

To evaluate the quality of a VCF file, different metrics are calculated using `granite qcVCF`. The software calculates both sample-based, as well as, family-based metrics.

The metrics currently available for sample are:

- variant types distribution
- base substitutions
- transition-transversion ratio

- heterozygosity ratio
- depth of coverage (GATK)
- depth of coverage (raw)

The metrics currently available for family are:

- mendelian errors in trio

1.2.2 Definitions

variant types distribution

Total number of variants classified by type as:

- **DEL**etion (*ACTG>A or ACTG>**)
- **INS**ertion (*A>ACTG or *>ACTG*)
- **S**ingle-**N**ucleotide **V**ariant (*A>T*)
- **M**ulti-**A**llelic **V**ariant (*A>T,C*)
- **M**ulti-**N**ucleotide **V**ariant (*AA>TT*)

base substitutions

Total number of SNVs classified by the type of substitution (e.g. C>T).

transition-transversion ratio

Ratio of transitions to transversions in SNVs. It is expected to be [2, 2.20] for WGS and [2.6, 3.3] for WES.

heterozygosity ratio

Ratio of heterozygous to alternate homozygous variants. It is expected to be [1.5, 2.5] for WGS analysis. Heterozygous and alternate homozygous sites are counted by variant type.

depth of coverage

Average depth of all variant sites called in the sample.

Depth of coverage (GATK) is calculated based on DP values as assigned by GATK. Depth of coverage (raw) is calculated based on raw read counts calculated directly from the bam file.

mendelian errors in trio

Variant sites in proband that are not consistent with mendelian inheritance rules based on parent genotypes. Mendelian errors are counted by variant type and classified based on genotype combinations in trio as:

Proband	Father	Mother	Type
0/1	0/0	0/0	de novo
0/1	1/1	1/1	error
1/1	0/0	(any)	error
1/1	(any)	0/0	error
1/1 | 0/1	./.	(any)	missing in parent
1/1 | 0/1	(any)	./.	missing in parent

1.3 BAM Quality Control

1.3.1 Overview

To evaluate the quality of a BAM file, different metrics are calculated using a custom script `bamqc.py`.

The metrics currently available are:

- mapping stats - total reads - reads w/ both mates mapped - reads w/ one mate mapped - reads w/ neither mate mapped
- read length
- coverage

1.3.2 Definitions

Mapping stats

The number of reads (not alignments) are counted as number unique read pairs, i.e. if a read pair is mapped to multiple locations, it is counted once.

Coverage

Coverage (=Depth of Coverage) is calculated as below:

```
{ (number of reads w/ both mates mapped) * (read length) * 2 + (number of reads w/␣
↪one mate mapped) * (read length) } / (effective genome size)
```

where effective genome size is the number of non-N bases in the genome.