PAC USER MANUAL

NAME

PAC – accurate allelic quantification at site and haplotype level

SYNOPSIS

```
nextflow run PAC/main.nf [options] --genome_version <genome
version> –reads <path to reads> –variants <path to variants> –
id <id> -profile <profile>
```

DESCRIPTION

Allele-specific expression (ASE) is the imbalanced expression of the two alleles of a gene. While many genes are expressed equally from both alleles, gene regulatory differences driven by genetic changes (i.e. regulatory variants) frequently cause the two alleles to be expressed at different levels, resulting in allele-specific expression patterns. The detection of ASE events relies on accurate alignment of RNA-sequencing reads, where challenges still remain. This pipeline has been created to adjust for computational biases associated with allelic counts. It comprises of the following steps:

1. Local phasing of genetic data using PHASER
2. Creation of parental genomes to align sequencing data to
3. Re-allocation of multimapping reads using RSEM
4. Selection of the best mapping for each read across the two parental genomes
5. Outputs haplotype and site level allelic counts

To run a test sample, run following commands:

```
load java

load singularity

git clone https://github.com/anna-saukkonen/PAC.git
```

```
path_to_nextflow/nextflow run PAC/main.nf --
genome_version GRCh37 --reads
"PAC/test/NA12890_merged_sample_0.005_{1,2}.fq.gz" --
variants
"PAC/test/NA12877_output.phased.downsampled.vcf.gz" --id
NA12877 -profile singularity
```

OPTIONS

Required:

--genome_version <genome version>

The available genomes are: GRCh37 or GRCh38.

--reads <path to reads>

The path to reads in within quotation marks. The reads need to be in the same directory with the following format: path_to_read_1.fq.gz and path_to_read_2.fq.gz. The options is called with: "path_to_reads_{1,2}.fq.gz".

--variants <path to variants>

The path to phased VCF file within quotation marks.

--id <id>

The sample ID.

-profile <profile>

The available options are: docker or singularity.

Optional:

-cpus

The number of cpus (as an integer). The default is 10.

-outdir

The name of the output file directory. The default is "/pac_results".

-N

An email address should the user want an email notification when the run is finished.

OUTPUT

PAC generates 5 output files:

haplotype level ASE calls:

1. ‘id’_gene_level_ae.txt

single nucleotide level ASE calls from PAC:

2. results_2genomes_‘id’.RSEM.STAR.SOFT.NOTRIM_baq.txt
3. results_2genomes_‘id’.RSEM.STAR.SOFT.NOTRIM.txt

single nucleotide level ASE calls based on standard single genome mapping for comparison:

4. results_1genome_‘id’.SOFT.NOTRIM_baq.txt
5. results_1genome_‘id’.SOFT.NOTRIM.txt

PREREQUISITE

Nextflow

The Nextflow can be downloaded with following command:

```
curl -fsSL get.nextflow.io | bash
```

Java

Java version 8 and above. You can check your java version with following command:

```
java -version
```

Docker or Singularity

The user needs a docker or singularity installed depending on which profile they use.