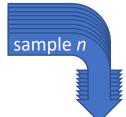


Healthy Code for Future Research

trim each fastq file with fastp

01_trim-fastq.py



• map each sample to ref.fa using bwa-mem

- add read group info while mapping
 - reads filtered with samtools
- record flagstats with samtools
- mapping coordinates recorded with bedtools

02_bwa-map_view_sort_index_flagstat.py

remove duplicates with picardtoolsrecord flagstats with samtools

03 mark build.py

• use GATK3's

RealignerTargetCreator and IndelRealigner to realign around indels

Once all sample.bam files have been religned ...

sample *n*

• Call SNPs across all samples for a given interval.bed using varscan

name

For each pool

sample *n*

sample *n*

Convert each batch.vcf to batch.txt

batch 00.bed

batch_01.b batch_02.b batch_03.b batch_05.b batch_06.b batch_07.b batch_08.b batch_09.b batch_10.b

start_varscan.py

Once all varscan jobs have finished ...

- Combine all batch.txt files and filter SNPs for GQ < 20, missing data > 25%, MAF
- Filter futher into files for INDELs, SNPs, SNPs in repeat regions, SNPs at putative paralog regions,

combine_varscan.py (filter_VariantsToTable.py)

Once combining has finished for each pool_name ...

- Get read stats (counts of reads at each stage of pipeline: raw -> mapped)
- Bundle files for transfer to local server (create rsync cmds, mkdir cmds)

(run manually) 98_get_read_stats.py
(run manually) 99_bundle_files_for_transfer.py