NISC Whole Genome Analysis Pipeline using Parabricks

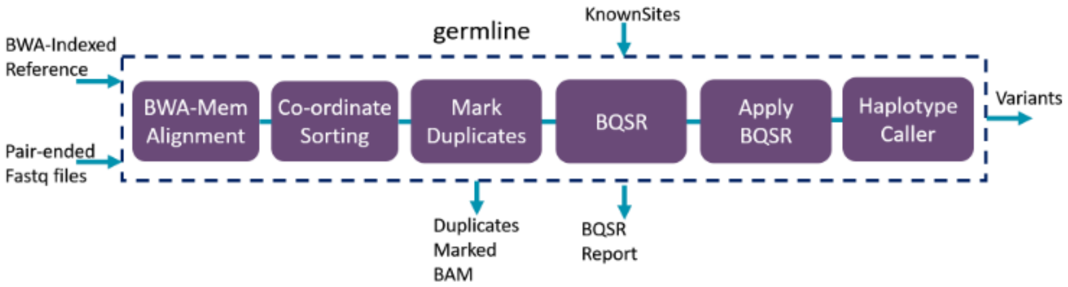
CONTENTS

1. Overview of Parabricks software
2. Reference genome
3. GATK Resource Bundle
4. Description of major steps in the GATK4 pipeline
5. NISC Deliverables
6. Additional information (gender, conversion to FASTQ)
7. PCR-Free Whole Genome Sequencing protocol
8. Link to commands, parameters and filters

Updated links 10/4/22

1. Overview of Parabricks software:

"Parabricks Human_Par Pipeline" is a GPU accelerated alignment and variant calling pipeline that uses the BWA and GATK algorithms under the hood.
https://docs.nvidia.com/clara/parabricks/v3.6.1/
https://docs.nvidia.com/clara/parabricks/v3.6.1/text/human_par_pipeline.html



2. Reference genome:

human_g1k_v37_decoy.fasta

3. GATK Resource Bundle:

B37

4. Description of major steps in the GATK4 pipeline:

The pipeline follows the GATK Best Practices for Variant calling and Filtration.
https://docs.nvidia.com/clara/parabricks/v3.6.1/text/human_par_pipeline.html

Sequence data for a sample is aligned to the "b37_decoy" reference using BWA (v0.7.15) as part of the pbrun human_par pipeline.

The software then sorts the data, marks duplicates, performs BQSR and runs the GATK Haplotype Caller.

NISC also generates filtered SNP+INDEL VCF files as per Broad’s recommendation.
https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants

Updated links 10/4/22

5. NISC DELIVERABLES:

BWA aligned sample bam
Sample gvcf
Filtered SNP and INDEL VCF (Autosomes only)

6. Additional Information:

- The gender is predicted using both read data and chip data.
  Females : X ploidy 2
  Y ploidy 1 (We return calls on Y so you have all the information)
  Males: X non-par ploidy 1
  X par ploidy 2
  Y ploidy 1

- Conversion of aligned BAM to FASTQ using Picard sam2fastq:
  https://broadinstitute.github.io/picard/command-line-overview.html#SamToFastq
  You will need to use the option --OUTPUT_PER_RG to produce FASTQ files per readgroup

7. PCR-Free Whole Genome Sequencing protocol

PCR-free libraries are generated from 1 microgram genomic DNA using the TruSeq® DNA PCR-Free HT Sample Preparation Kit (Illumina). The median insert sizes are approximately 400 bp. Libraries are tagged with unique dual index DNA barcodes to allow pooling of libraries and minimize the impact of barcode hopping. Libraries are pooled for sequencing on the NovaSeq 6000 (Illumina) to obtain at least 300 million 151-base read pairs per individual library.

8. Commands, parameters and filters:

https://docs.nvidia.com/clara/parabricks/v3.6.1/text/human_par_pipeline.html

Updated links 10/4/22