SNPTools integrative variant calling and genotype/haplotype imputation pipeline

v 1.0 June 28,2012

Authors: Yi Wang, Jin Yu, James Lu, and Fuli Yu

Human Genome Sequencing Center (HGSC) at Baylor College of Medicine (BCM)

Houston, TX, USA

Contact: Jin Yu (jy2@bcm.edu), Fuli Yu (fyu@bcm.edu)

1. Introduction

SNPTools is a suite of tools that enables integrative SNP analysis in next generation sequencing data with large cohorts. It not only calls SNP in a population with high sensitivity and accuracy, but also employs a novel imputation engine to achieve highly accurate genotype calls in an efficient way.

2. System Requirements and Installation

- Unix-like operation system
- C++ compiler and Make must be installed

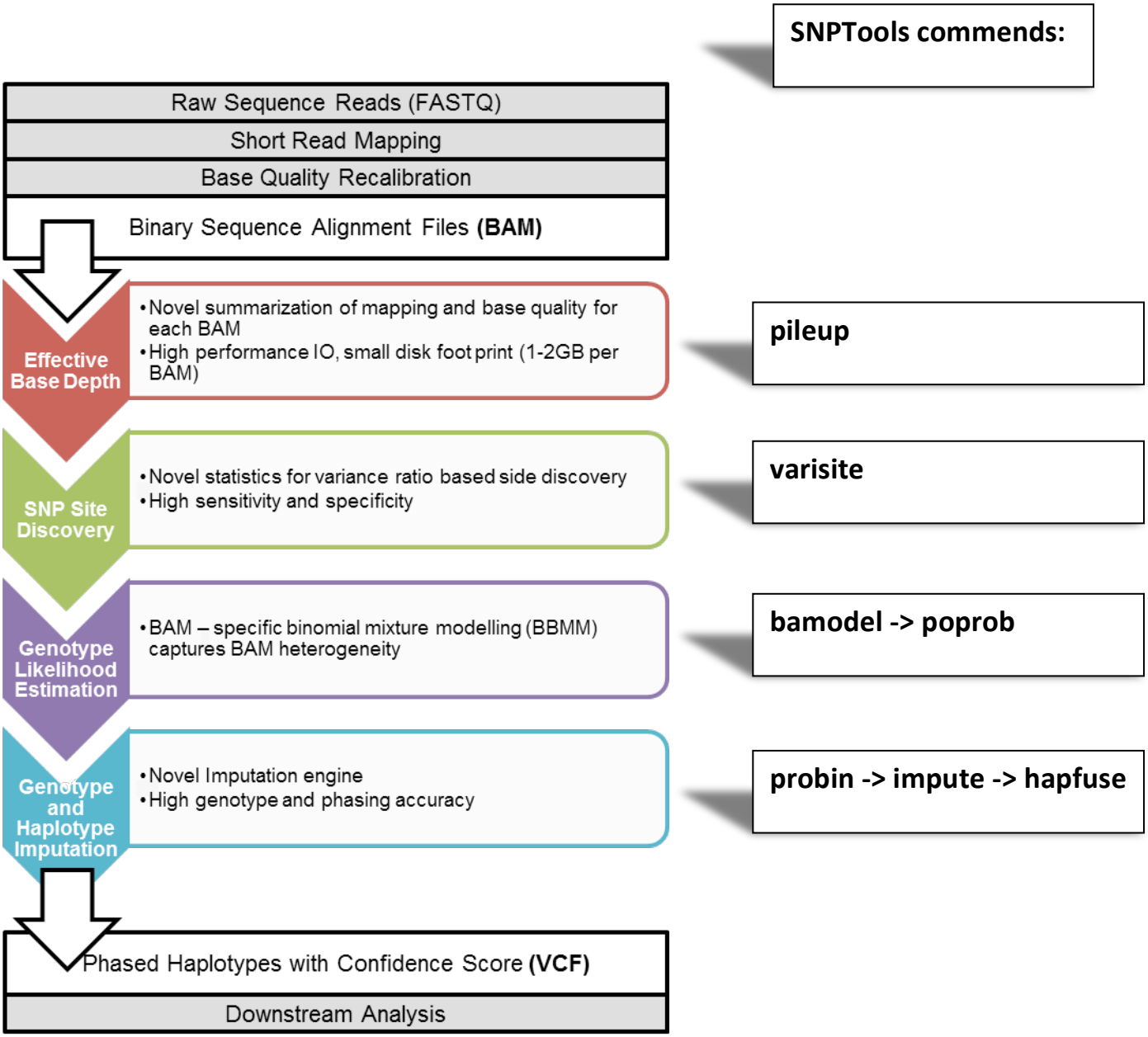
Typing "make" will compile the codes and generate all the executive binary files of SNPTools.

3. Preprocessing

Although SNPTools will function on any sorted BAM file, there are number of preprocessing steps that we recommend for the highest quality results.

- Mark or remove PCR duplicates using PicardMarkDuplicates (http://picard.sourceforge.net/index.shtml) or a similar tool.
- Base quality recalibration (not recommended for SOLiD data) using GATK (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit) or a similar tool.
- Locally realign around likely indels and using GATK or SMRA (http://sourceforge.net/projects/srma/).

4. Workflow Summary



5. Usage

5.1 pileup

pileup is the tool to extract read depth from BAM. After running this commend, it will create an EBD file and an EBI file for each BAM.

```
pileup [1.bam 2.bam ...]
```

5.2 varisite

varisite is a SNP calling tool of population NGS data. It takes sequencing data in EBD format as input and out the candidate SNP sites in vcf format. The list of EBD files are listed in a text file ebd.list.

```
varisite [options] <ebd.list chr chr.fa>
```

-i consider indel allele (false)

-s <FLT> significance cutoff (1.5)

-v <site.vcf> only scoring on give sites

-g <bit_mask> group files by fields specified with bit_mask (0)

-l <FLT> variance stabilizer (1)

5.3 bamodel

bamodel is the tool to calculate the genotype likelihood of each bam at given sites using a BAM by specific binomial mixture model (BBMM). It will output an individual likelihood file in binary format of each BAM.

```
bamodel [options] <out.raw site.vcf in1.bam> [in2.bam in3.bam...]
```

```
-a <FLOAT>      P(read=ref|geno=alt/alt) (0.010)
-h <FLOAT>      P(read=ref|geno=ref/alt) (0.500)
-r <FLOAT>      P(read=ref|geno=ref/ref) (0.995)
-p <FLOAT>      precision of beta distribution prior (100)
```

5.4 poprob

poprob is a data formater tool to merge and transpose individual likelihood files to a single population likelihood file

```
poprob [options] <site.vcf raws.list out.prob>
```

```
      -b <INT>    buffer size in MB (1024)
```

5.5 impute

impute is the genotype/haplotype imputation engine. It takes chunked population genotype likelihood files as input and output the imputed genotype and haplotype of each chunk of genome.

```
impute [options] 1.bin 2.bin ...
```

```
       -d <density>   relative SNP density to Sanger sequencing (1)
       -b <burn>      burn-in generations (64)
       -l <file>      list of input files
       -m <mcmc>      sampling generations (192)
       -n <mcmc>      nested MH sampler iteration (1024)
       -t <thread>    number of threads (0=MAX)
       -v <vcf>       integrate known genotype in VCF format
```

5.6 hapfuse

hapfuse is the tool to merge the chunked haplotypes result files to a single vcf file of each chromosome.

```
hapfuse <out.vcf dir>
```

6. References to SNPTools

SNPTools contributed to the Phase1 main project of the 1000 Genomes Project and is currently contributing to Phase 2 of the main project. It is referenced by Wang, Y. etc. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. American Journal of Human Genetics (in review, 2012)