Genomes on the Cloud : GotCloud

IBG Workshop May 6, 2013

University of Michigan
Center for Statistical Genetics

Mary Kate Wing
Goo Jun

What is GotCloud?
Mapping and Variant Calling Pipeline

- Connects sequence analysis tools together
  - Alignment, quality control, variant calling
- Divides large jobs into many small pieces
  - Simplifies running on clusters, re-starting after failure
- Can analyze many samples together
- Can run on Amazon Cloud

Running on Amazon

- Estimates for one 92x Exome sample
  - medium instance
  - 3.75GB memory, 2 compute units, 410GB storage

Description	Size	Time	Cost*
Upload Reference Files	6 G	1 hr	$0.13
Copy fastqs from 1000G Amazon S3	7.5 G	1 hr	$0.13
Alignment Pipeline 2 single-end, 2 paired-end (6 Total Fastqs)	7.5 G	25 hrs	$3.25
Variant Calling Pipeline (without LD-aware genotyping)	13 G	5 hrs	$0.65

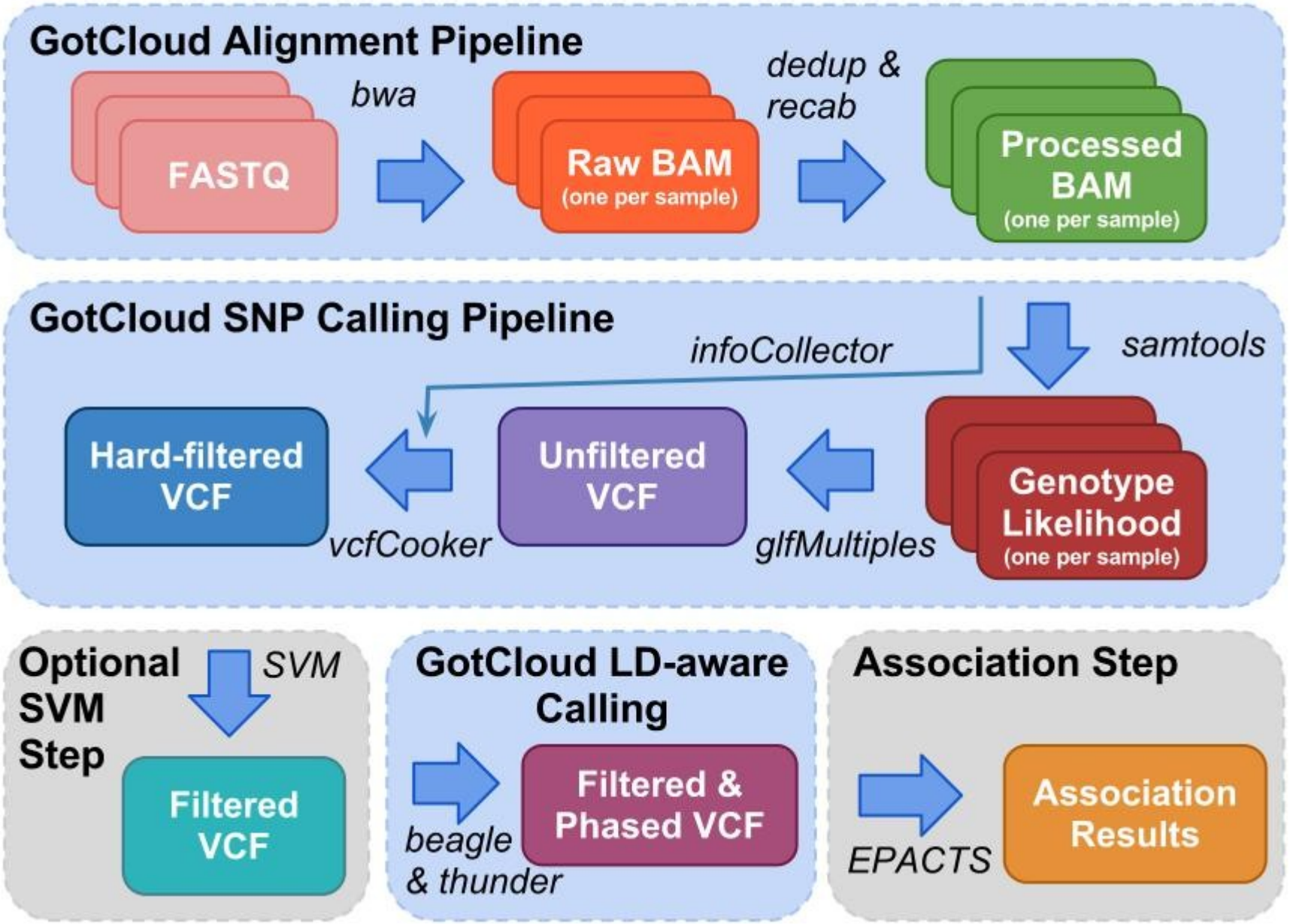
*EC2 cost can fluctuate. This cost estimate is for $0.13/hr

Sequence Data File Format : FASTQ

- “raw” output of a sequencing run
  - Sequence reads, each with a **label**, sequence, and base qualities
  - Far too much information to be human readable
- Reads are often “paired”

@TYPICALLY_CRYPTIC_READ_NAME
GAAATTCATCTGTCCTCAGACACAGG
+
BGGG?GEGGGGGGFFGGGG:GFFG:EEB
@ANOTHER_CRYPTIC_READ_NAME
AGAGTCTCACTCTGTCCCTCAGGCTGG
+
GEGGGGGGGEGGGGGGGGFEFGEGEG?

Next Generation Sequence Data Processing Pipeline



First Step: Alignment

- Finds most likely genomic location for each read
- Merge together multiple sequencing runs for a sample
- Cleanup the initial alignment
  - Remove Duplicate Reads
  - Ensure base quality estimates are accurate
- Generate visual summaries for inspections
- Check for sample contamination (more tomorrow)

First Bit of Output: BAM Files

- Binary implementation of Sequence Alignment/Map (SAM) format
  - http://genome.sph.umich.edu/wiki/BAM
- Records best position for each read
  - Starting point for variant discovery and genotyping
- Not really for human consumption
- You can get a summary of alignments using samtools tview

```
samtools tview \
/opt/gotcloudExample/bams/HG00254.bam \
/opt/gotcloudExample/chr20Ref/human_g1k_v37_chr20.fa
```

  - Type “g”, then: “20:42900000”, Type “q” to quit

Second Step: Variant Discovery

- Starts from aligned data in BAM files
- Generates list of variant sites and genotypes
  - Optionally, uses haplotype information to refine genotypes
- This step includes filtering out low quality reads, modeling overlapping reads, adjusting alignments, flagging poor quality variants...
- Jobs can be divided into many small pieces before result is merged back together

Second Step Output: VCF Files

- Stores a list of variant sites, alleles, and positions
  - Can also store additional information about each variant, but format is cryptic
  - Usually has a header with clues about the content
- Each site, and often each genotype, has a quality
- Can optionally store individual genotypes
- http://tinyurl.com/VCF4-1

Sample VCF Files

***** Only with variant information *****

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	42900	rs123	G	T	100	PASS	DP=213;MQ=56;AC=14;AF=0.142
20	42901	.	A	T	100	PASS	DP=236;MQ=57;AC=7;AF=0.066

***** Also with genotypes *****

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
20	42900	rs123	G	T	100	PASS	DP=213	GT:DP	0/0:100	1/1:113
20	42901	.	A	T	100	PASS	DP=236	GT:DP	0/1:90	1/1:146

Todays Tutorial

- Run alignment pipeline to map two samples
  - Review summaries of alignment
- Generate variant calls for 60 indivduals from 1000 Genomes Project
  - Generate list of variants in VCF format
- Input files
  - Input files in /opt/gotcloudExample
  - Binary files already installed
- To conserve time and disk-space, analysis focuses on a small region of chromosome 20, 42900000 - 43200000

Getting Things Ready

- Start a UNIX terminal
- Test data already installed on each machine in /opt/gotcloudExample, no need to copy
- Review input, index, and configuration files
  - No editing is required for the tutorial

Run the Alignment Pipeline

- Type the following in your terminal to run:

gotcloud align --conf /opt/gotcloudExample/GBR2align.conf --outdir ~/gotcloudTutorial

- This runs the whole alignment pipeline

FASTQ Index File

MERGE_NAME	FASTQ1	FASTQ2	RGID	SAMPLE	LIBRARY	CENTER	PLATFORM
SM1	/data/SM1_RG1_1.fastq	/data/SM1_RG1_2.fastq	RG1	SM1	lib1	WUGSC	ILLUMINA
SM1	/data/SM1_RG1.fastq	.	RG1	SM1	lib1	WUGSC	ILLUMINA
SM1	/data/SM1_RG2_1.fastq	/data/SM1_RG2_2.fastq	RG2	SM1	lib1	WUGSC	ILLUMINA
SM2	/data/SM2_RG3.fastq	.	RG3	SM2	lib2	SC	ILLUMINA

To see the index file for your analysis, use:

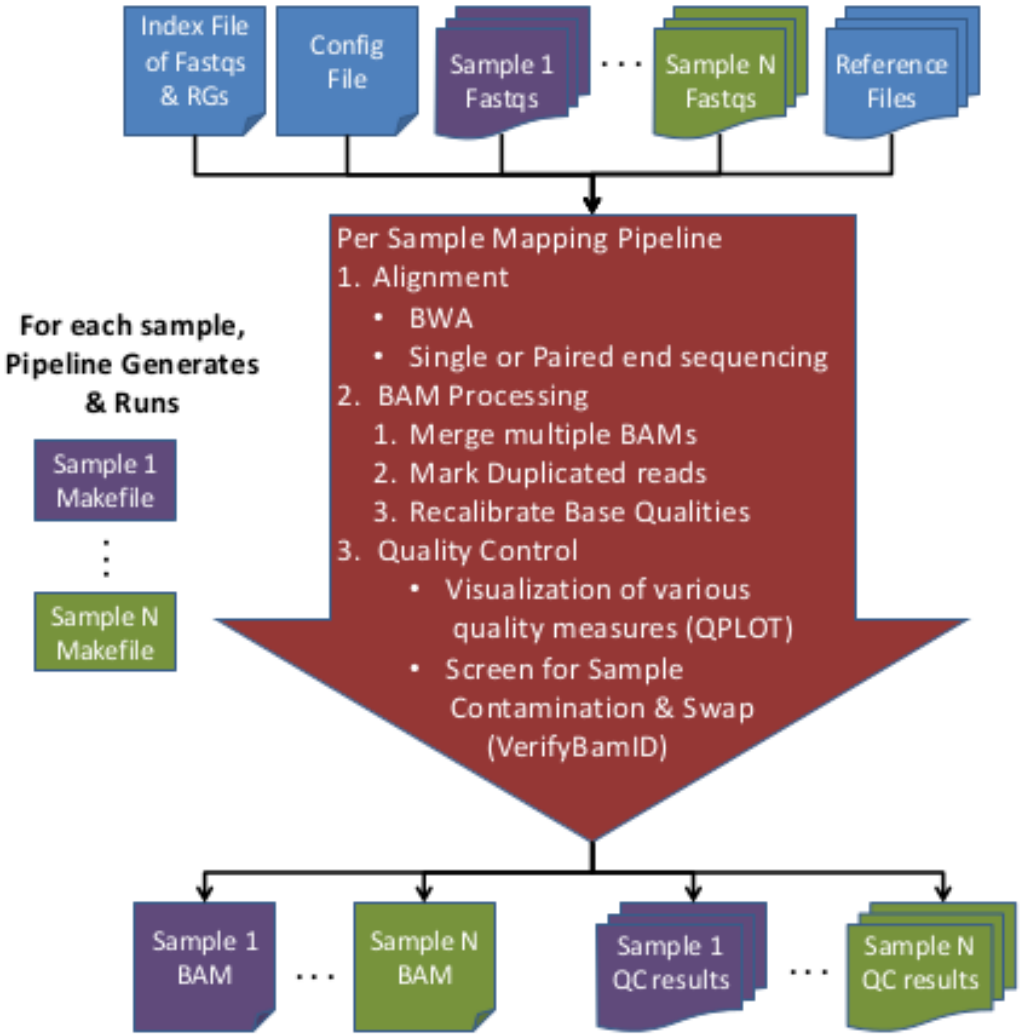
less -S /opt/gotcloudExample/GBR2fastq.index

Tutorial Configuration File

```
INDEX_FILE = /opt/gotcloudExample/GBR2fastq.index
#############
# References
REF_DIR = /opt/gotcloudExample/chr20Ref
AS = NCBI37
FA_REF = $(REF_DIR)/human_g1k_v37_chr20.fa
DBSNP_VCF =  $(REF_DIR)/dbsnp135_chr20.vcf.gz
HM3_VCF = $(REF_DIR)/hapmap_3.3.b37.sites.chr20.vcf.gz
```

- Series of KEY=VALUE pairs
- Specifies input files, reference genome, and related info
- Note: Tutorial uses chromosome 20 only references to speed processing

Alignment Pipeline



Alignment Pipeline Output

- Takes about 1-2 minutes
- On success generates aligned data for each sample

  Processing finished in nn secs with no errors reported

- Output Files:

  ls ~/gotcloudTutorial/bams/*bam

  ls ~/gotcloudTutorial/QCFiles/*qplot*

  ls ~/gotcloudTutorial/QCFiles/*genoCheck*

Alignment Pipeline: Reviewing Output

- Review alignment

  samtools tview \
  ~/gotcloudTutorial/bams/HG00096.recal.bam \
  /opt/gotcloudExample/chr20Ref/human_g1k_v37_chr20.fa

  - Type “g”, then: “20:42900000”
  - Type “q” to quit
- Look at contamination estimates

  more ~/gotcloudTutorial/QCFiles/*.selfSM

  - Note: HG00100 has high FREEMIX, but that is due to the small region being analyzed

Alignment Pipeline: Reviewing Output

- Review Alignment Statistics

  more ~/gotcloudTutorial/QCFiles/*.stats

- Review pdfs of plot results

  okular ~/gotcloudTutorial/QCFiles/HG000096.qplot.pdf

  - Note: recalibrated file's phred score looks bad, but this is due to the small region analyzed
    - See the following link for whole genome plots

http://genome.sph.umich.edu/w/images/f/f2/HG00096.wg.qplot.pdf

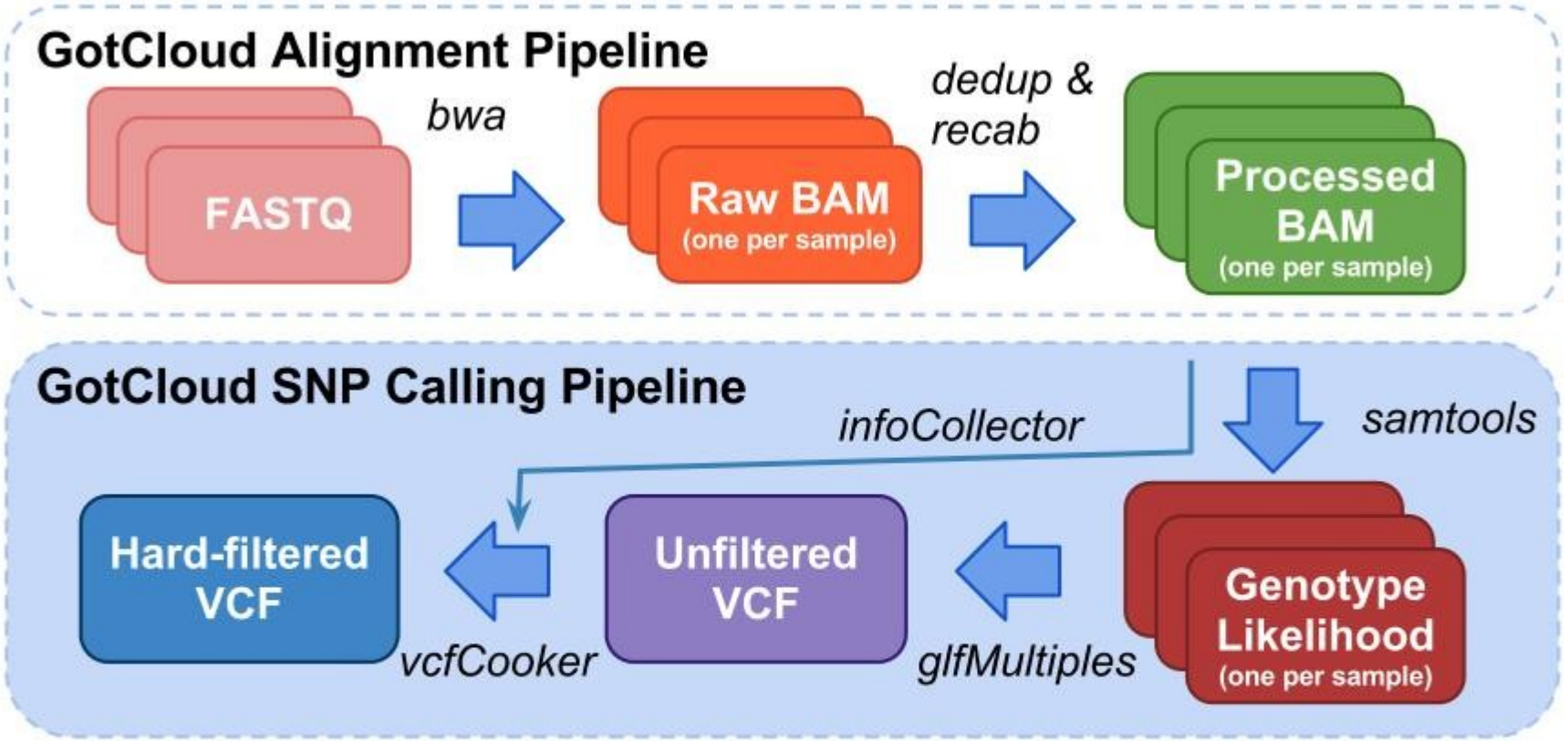
Run the SNP Calling Pipeline

- Type the following in your terminal to run:

  gotcloud snpcall --conf /opt/gotcloudExample/GBR60vc.conf \

  --outdir ~/gotcloudTutorial --numjobs 2 --region 20:42900000-43200000

- This runs the snp calling pipeline



Variant Calling Pipeline

- Now that samples are aligned, we are ready to do variant calling
- Run 60 low pass samples others
  - Output of alignment pipeline could be used as input for the variant calling pipeline

Tutorial BAM Index File

- Look at the BAM Index File:

  more /opt/gotcloudExample/GBR60bam.index

HG00096	GBR	bams/HG00096.bam
HG00100	GBR	bams/HG00100.bam
HG00103	GBR	bams/HG00103.bam
HG00106	GBR	bams/HG00106.bam
HG00108	GBR	bams/HG00108.bam
HG00111	GBR	bams/HG00111.bam

- Only one BAM file per sample in this case
  - Would specify more by adding tab delmited columns

Variant Calling Configuration File

```
CHRS = 20
BAM_INDEX = /opt/gotcloudExample/GBR60bam.index
############
# References
REF_ROOT = /opt/gotcloudExample/chr20Ref
#
REF = $(REF_ROOT)/human_g1k_v37_chr20.fa
INDEL_PREFIX = $(REF_ROOT)/1kg.pilot_release.merged.indels.sites.hg19
DBSNP_VCF =  $(REF_ROOT)/dbsnp135_chr20.vcf.gz
HM3_VCF =  $(REF_ROOT)/hapmap_3.3.b37.sites.chr20.vcf.gz

# Update Thunder settings to run faster for the tutorial:
# Run 10 rounds instead of 30 (-r 10)
# Run without --compact to run faster but use more memory
#   This works for the small tutorial set
THUNDER = $(UMAKE_ROOT)/bin/thunderVCF -r 10 --phase --dosage --inputPhased
```

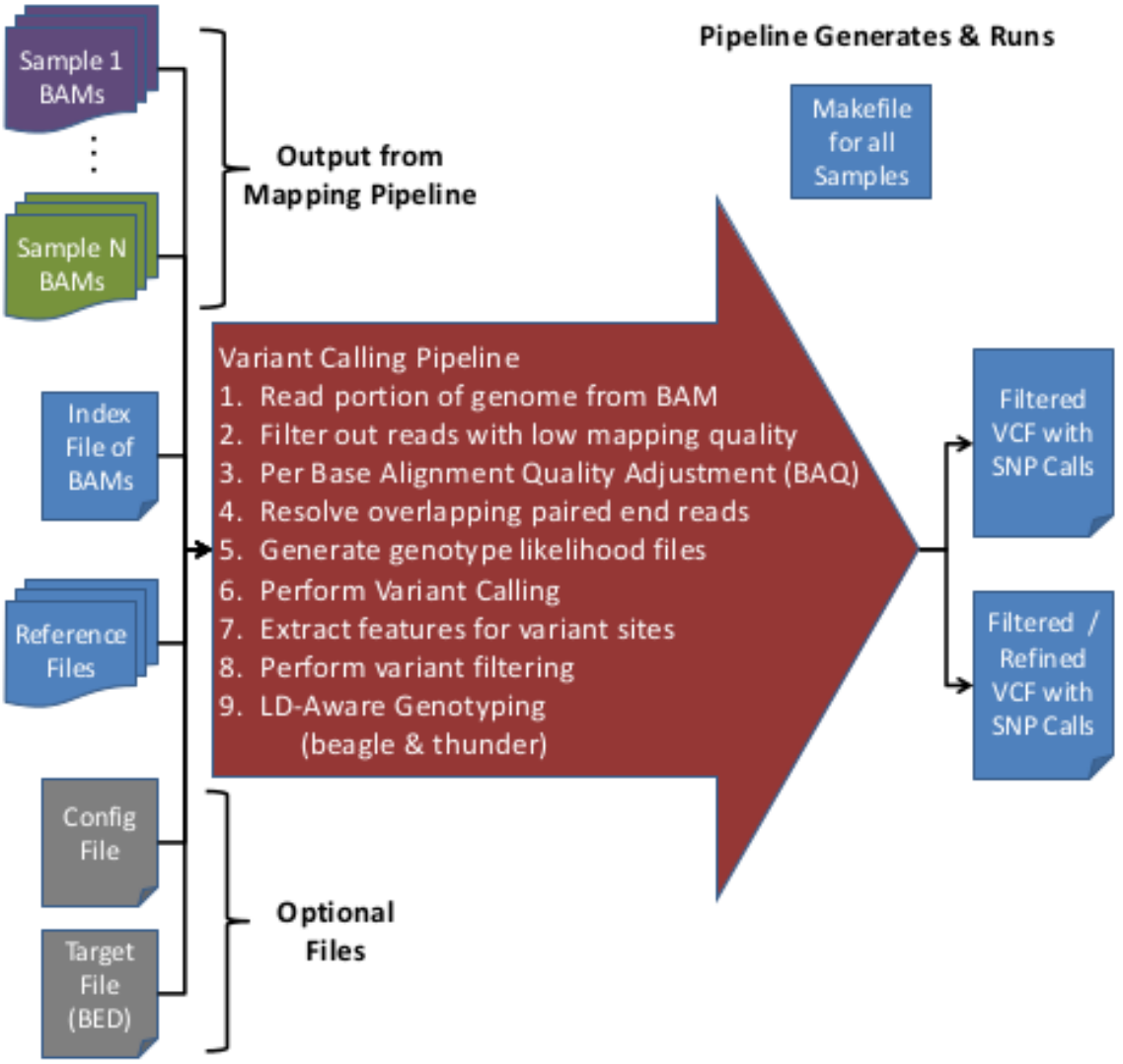
more /opt/gotcloudExample/GBR60vc.conf

- Note: The Tutorial uses chromosome 20 only reference files to speed processing

If you were analyzing your own data…

- Update BAM_INDEX file
- Download whole genome reference files and indicate their location
- Tweak cluster configuration and number of jobs to match your cluster
- Do not use the modified Thunder Settings
- Add configuration for Targeted regions if applicable

Variant Calling Pipeline



SNP Calling Pipeline Output

- Takes about 3-5 minutes
- On success, generates a merged VCF file:

  Commands finished in nnn secs with no errors reported

- Output Files:

  ls ~/gotcloudTutorial/split/chr20/chr20.filtered*

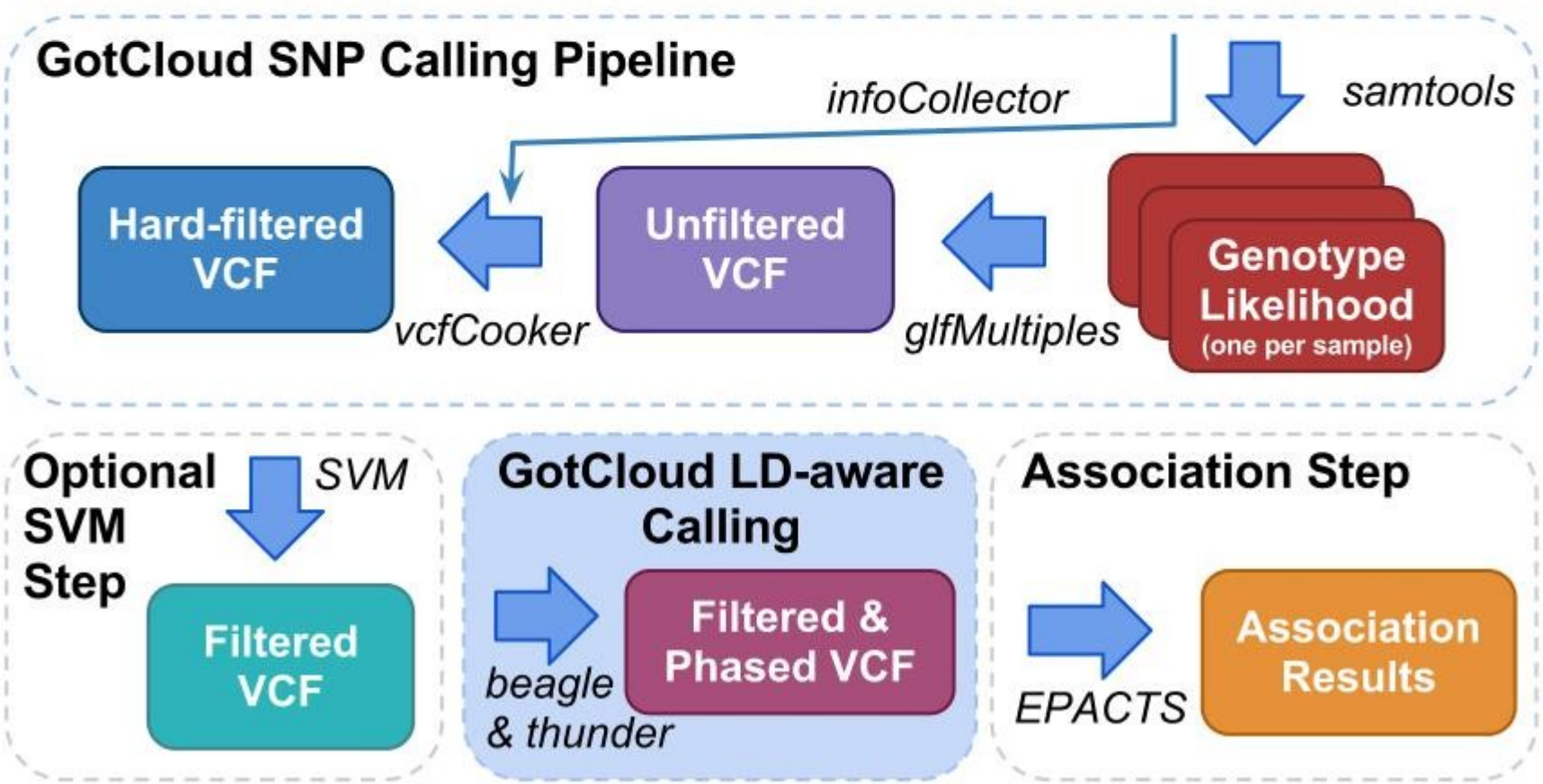
  ls ~/gotcloudTutorial/vcfs/chr20/chr20.filtered*

- Key outputs are :
  - split/chr20/chr20.filtered.PASS.vcf.gz – high quality sites
  - vcfs/chr20/chr20.filtered.sites.vcf.summary – filtering summary

Run the LD-aware Genotype Refinement Pipeline

- Type the following in your terminal to run:

```
gotcloud ldrefine --conf /opt/gotcloudExample/GBR60vc.conf \
--outdir ~/gotcloudTutorial --numjobs 2
```



LD-Aware Calling Pipeline Output

- Takes about 2-3 minutes
- On Success:

  Commands finished in nnn secs with no errors reported

- Final output file with updated genotypes:

  ls ~/gotcloudTutorial/thunder/chr20/GBR/thunder/chr20.filtered.PASS.beagled.GBR.thunder.vcf.gz

GotCloud plugs into EPACTS for Association Analysis

We'll skip that today.

Run the Association Analysis Pipeline (EPACTS)

- Efficient and Parallelizable Association Container Toolbox (EPACTS)
- Type the following in your terminal to run:

```
epacts single --vcf ~/gotcloudTutorial/vcfs/chr20/chr20.filtered.vcf.gz \
--ped /opt/gotcloludExample/test.GBR60.ped --out ~/gotcloudTutorial/epacts/epacts \
--test q.linear --run 1 --top 1 --chr 20
```

EPACTS Output

- Takes about 1-3 minutes
- On Success:

  Commands finished in nnn secs with no errors reported
- Output Files:
  - ls ~/gotcloudTutorial/epacts/*
- To see the top associated variants, you can run
  - less ~/gotcloudTutorial/epacts/epacts.top5000
- To see the locus-zoom like plot:
  - xpdf ~/gotcloudTutorial/epacts/epacst.zoom.20.42987877.pdf

EPACTS Output PDF