Practical Introduction

Variant Calling and Filtering for SNPs

May 19, 2015

Mary Kate Wing
Hyun Min Kang

Goals of This Session

- Learn basics of Variant Call Format (VCF)
- Aligned sequences -> filtered snp calls
  - Many methods/pipelines, we cover 1
- Examine variants at particular genomic positions
- Evaluate quality of SNP calls

Variant Call Format (VCF)

- Describes variant positions
  - http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
- Header
  - Each line starts with #
- Records
  - One for each variant position
  - Describes variant
  - Optional per sample genotype information

Variant Call Format: Header

```
##fileformat=VCFv4.1
##filedate=20140615
##source=glfMultiples
##minDepth=1
##maxDepth=10000000
##minMapQuality=0
##minPosterior=0.5000
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth at Site">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root Mean Squared Mapping Quality">
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Coverage">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Number of Alleles in Samples with Coverage">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Counts in Samples with Coverage">
##INFO=<ID=AF,Number=.,Type=Float,Description="Alternate Allele Frequencies">
##INFO=<ID=MQ30,Number=1,Type=Float,Description="Fraction of bases with mapQ<=30">
##FILTER=<ID=mq0,Description="Mapping Quality Below 0">
##FILTER=<ID=dp1,Description="Total Read Depth Below 1">
##FILTER=<ID=DP10000000,Description="Total Read Depth Above 10000000">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Most Likely Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Call Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=PL,Number=.,Type=Integer,Description="Genotype Likelihoods for Genotypes in Phred Scale, fo
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00551 HG00553 HG00554 HG00637
```

Description of INFO, FILTER, & FORMAT fields

Description of the records fields

Order of per samples genotypes

Variant Call Format: Records

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00551	HG00553
22	35999938		.	1) A	G	100	PASS	DP=127;MQ=59;NS=53;AN=10		
22	36000547		.	2) A	G	100	PASS	DP=485;MQ=59;NS=62;AN=12		
22	36000711		.	3) G	T	24	PASS	DP=376;MQ=59;NS=61;AN=12		
22	36707786		.	4) A	G,C	100	PASS	DP=373;MQ=59;NS=59;AN=11		
				A	B					

SNPs A: Reference B: Alternate

1) Alternate G 2) Alternate G

3) Alternate T 4) 2 Alternates bases: G & C

				A	B			
22	16123409	.	1)	G	GA	21	PASS	AC=1;AF=0.0
22	16136754	.	2)	TG	T	26	PASS	AC=2;AF=0.0
22	16139950	.	3)	G	GA	19	PASS	AC=88;AF=0.
22	16140022	.	4)	AAAGG	A	100	PASS	AC=40;AF=0.

INDELs A: Reference B: Alternate

1) Insertion of A 2) Deletion of G

3) Insertion of A 4) Deletion of AAGG

Variant Call Format: Records

This sample is Homozygous Alt for this variant

This sample is Heterozygous for this variant

This sample is Homozygous Ref

```
GT:DP:GQ:PL     1/1:10:29:168,30,0     0/1:7:19:34,0,18        0/1:8:27:82,0
GT:DP:GQ:PL     1/1:10:29:150,30,0     0/1:7:25:34,0,24        0/1:8:20:83,0
GT:DP:GQ:PL     0/1:9:22:22,0,69       0/1:3:9:9,0,37  0/0:1:5:0,3,18  0/0:2
```

```
GT:DP:GQ:PL     1/1:7:15:73,21,0,73,21,73     2/2:1:5:23,23,23,3,3
```

This sample is Homozygous Alt1 for this variant

This sample is Homozygous Alt2 for this variant

Variant Call Format (VCF)

- It's a large file, how do I look at certain variants?
  - tabix
    - http://samtools.sourceforge.net/tabix.shtml
    - Generate tabix index (.tbi) file:
      - tabix -p vcf file.vcf.gz
    - View region:
      - tabix file.vcf.gz CHR:START-END

High Quality Variant Calls from BAMs

- Many tools & best practices to choose from
- Our solution:

Genomes on the Cloud (GotCloud)

  - Sequence analysis pipelines
    - You don't need to know the details of individual components
    - Automates steps for you

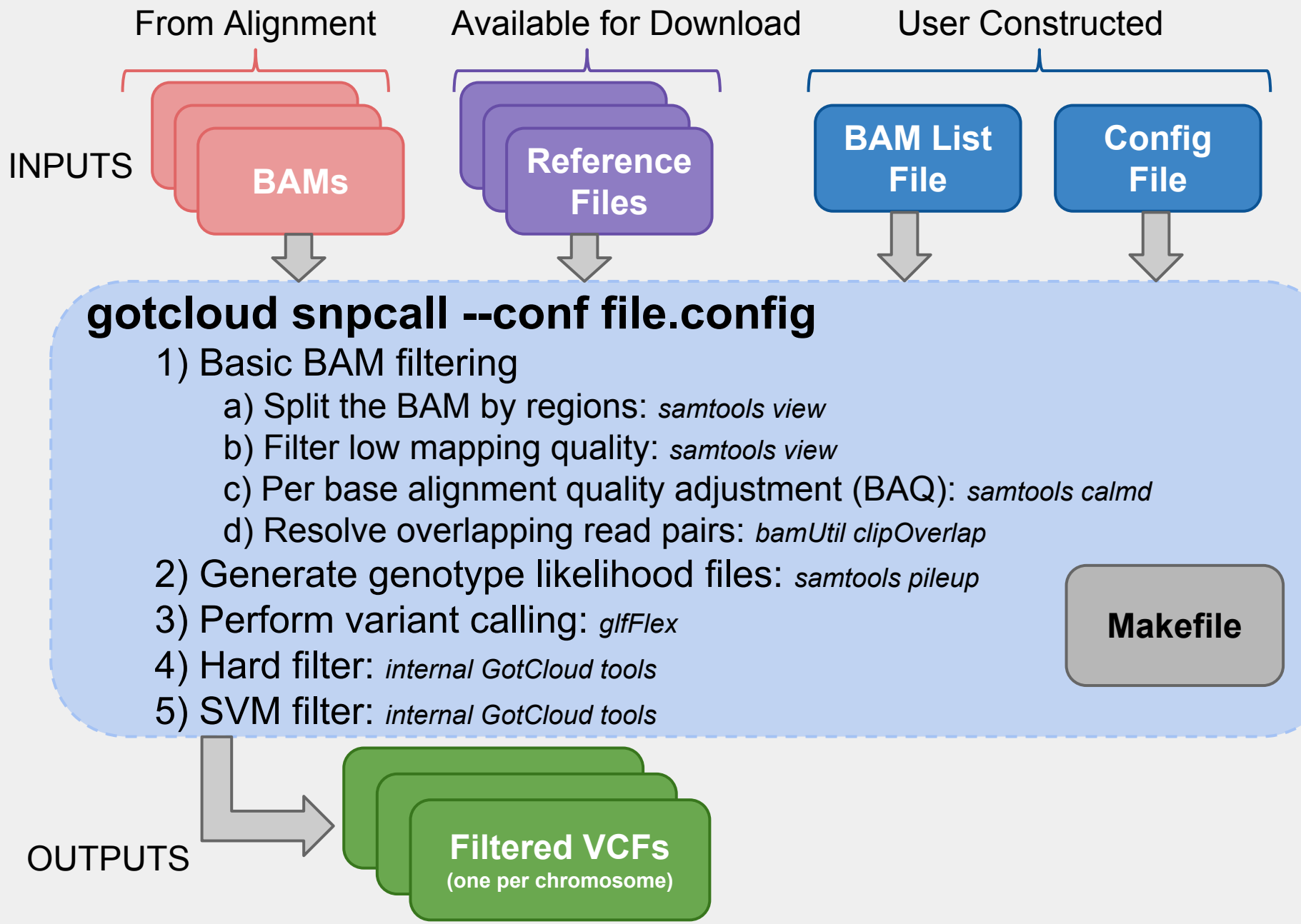
Why GotCloud snpcall?

- Robust parallelization
  - Automatically partitions **chromosomes by regions**
  - Takes advantage of clusters
    - Supports MOSIX, slurm, SGE, pbs
    - Can setup a cluster on Amazon
  - via GNU make
    - Reliable and fault-tolerant
    - Restart where it stopped
- Analyzes many samples together
- Easy to add new samples to your study

Studies using GotCloud snpcall

Study	Genome	Exome	N	Populations	# SNPs
1000 Genomes	~6x	~40x	2,535	Many	69.1M
Type 2 Diabetes	~5x	~80x	2,850	Europeans	26.7M
Exome Sequencing Project	.	~80x	6,916	EUR+AFR	1.92M
Sardinian Sequencing	~4x	.	3,520	Sardinians	23.1M
Bipolar Sequencing	~12x	.	2,825	Europeans	43.7N
Nephrotic Syndrome	~4x	.	464	Many	25.6M
Age-related Macular Degeneration	~6x	.	3,000	Europeans	36.2M
HUNT	~4x	.	1,200	Norwegians	23.0M

GotCloud SnpCall Pipeline Overview



Reference Files

- GotCloud snpcall uses:
  - Reference genome FASTA file
    - To identify differences (SNPs) between bases in sequence reads & the reference positions they mapped to
- VCF files
  - indel - contains known insertions & deletions to help with filtering
  - omni - used as likely true positives for SVM filtering
  - hapmap - used as likely true positives for SVM filtering and for generating summary statistics
  - dbsnp - used for generating summary statistics

User Created Input: BAM List File

- Points GotCloud to the BAMs
  - Alignment pipeline generates for you
  - For our tutorial: update it to include more BAMs
- Tab delimited

1) Sample name
one row per sample

2 .. N) BAM - typically only 1 BAM for sample,
but if more than one, separate with tabs

```
HG00641	/net/seqshop-server/home/mktrost/out/bams/HG00641.recal.bam
HG00640	/net/seqshop-server/home/mktrost/out/bams/HG00640.recal.bam
HG00551	/net/seqshop-server/home/mktrost/out/bams/HG00551.recal.bam
HG00553	/net/seqshop-server/home/mktrost/out/bams/HG00553.recal.bam
```

User Constructed Input: GotCloud Configuration

```
############ ← #'s are comments
# References
REF_DIR = ref22
REF = $(REF_DIR)/human.g1k.v37.chr22.fa
DBSNP_VCF =  $(REF_DIR)/dbsnp_135.b37.chr22.vcf.gz
HM3_VCF = $(REF_DIR)/hapmap_3.3.b37.sites.chr22.vcf.gz
INDEL_PREFIX = $(REF_DIR)/1kg.pilot_release.merged.indels.sites.hg19
OMNI_VCF = $(REF_DIR)/1000G_omni2.5.b37.sites.PASS.chr22.vcf.gz

######### ALIGNMENT ########
MAP_TYPE = BWA_MEM
FASTQ_LIST = fastq.list

######### Variant Calling ########
CHRS = 22
```

Use $(KEY) to refer to other KEYs

Path to chr22 reference files

Use bwa mem instead of just regular BWA

Path to fastq index file

For snpcall & indel -> chr22 only

User Constructed Input: GotCloud Configuration

```
######### THUNDER ########
# Update so it will run faster for the tutorial
#   * 10 rounds instead of 30 (-r 10)
#   * without --compact option
#   Runs faster, but uses more memory, but not a lot for the small example
THUNDER = $(BIN_DIR)/thunderVCF -r 10 --phase --dosage --inputPhased $(THUNDER_STATES)
```

Thunder Settings to speed up LD Refinement Pipeline for the tutorial

```
##############################
## GenomeSTRIP
##############################
GENOMESTRIP_MASK_FASTA = $(REF_DIR)/human_g1k_v37.chr22.mask.100.fasta
GENOMESTRIP_PLOIDY_MAP = $(REF_DIR)/humgen_g1k_v37_ploidy.chr22.map
```

Structural Variation Pipeline Settings

What will I need to configure in GotCloud for my own research?

- Exome/Targeted set in your configuration:

```
# When all individuals have the same target
UNIFORM_TARGET_BED = path/to/file.bed

# When each individual has different targets
# Each line of file.txt contains [SM_ID] [TARGET_BED]
MULTIPLE_TARGET_MAP = path/to/file.txt

# Extend target by given # of bases
OFFSET_OFF_TARGET = 0

# If a single chromosome is too small for SVM,
# set this to run SVM on all chromosomes combined.
# Only for very small targeted projects.
# Exome does not require this.
WGS_SVM = TRUE
```

What will I need to configure in GotCloud for my own research?

- Cluster support
  - Via configuration
    - BATCH_TYPE =
      - mosix, pbs, slurm, pbs, sge, slurmi, sgei
    - BATCH_OPTS =
      - Set to any options you would normally pass to your cluster
  - Via command line
    - --batchtype & --batchopts

How good are the results?

${OUT}/vcfs/chr*/chr*.filtered.sites.vcf.summary

FILTER	#SNPs	#dbSNP	%dbSNP	%CpG Known	%CpG Novel	%Known Ts/Tv	%Novel Ts/Tv	%nCpG-K Ts/Tv	%nCpG-N Ts/Tv	%HM3 sens	%HM3 /SNP
INDEL5	56	50	89.3	10.0	0.0	1.78	1.00	1.50	1.00	0.005	1.786
INDEL5;SVM	9	9	100.0	0.0	NA	0.80	NA	0.80	NA	0.000	0.000
PASS	3870	3741	96.7	21.9	17.1	2.36	2.23	1.94	1.82	2.325	12.403
SVM	129	112	86.8	16.1	17.6	3.31	1.83	2.92	1.80	0.000	0.000

FILTER	#SNPs	#dbSNP	%dbSNP	%CpG Known	%CpG Novel	%Known Ts/Tv	%Novel Ts/Tv	%nCpG-K Ts/Tv	%nCpG-N Ts/Tv	%HM3 sens	%HM3 /SNP
INDEL5	65	59	90.8	8.5	0.0	1.57	1.00	1.35	1.00	0.005	1.538
PASS	3870	3741	96.7	21.9	17.1	2.36	2.23	1.94	1.82	2.325	12.403
SVM	138	121	87.7	14.9	17.6	2.90	1.83	2.55	1.80	0.000	0.000

PASS	3870	3741	96.7	21.9	17.1	2.36	2.23	1.94	1.82	2.325	12.403
FAIL	194	171	88.1	13.5	13.0	2.49	1.56	2.15	1.50	0.005	0.515
TOTAL	4064	3912	96.3	21.5	16.4	2.37	2.10	1.95	1.76	2.330	11.836

MultiAllele Ref/Alt	1
Repeated Positions	0
TOTAL SKIPPED	1

Genotype Refinement

- After snpcall, we run genotype refinement
  - improves the genotypes - higher quality
  - Beagle & thunder
- Outputs are VCFs
  - thunder breaks up by population

Try it yourself

http://genome.sph.umich.edu/wiki/SeqShop:_Variant_Calling_and_Filtering_for_SNPs_Practical