

Variant calling

What are genetic variants

- **SNP (single-nucleotide polymorphism)**
 - a mutation or substitution of a single nucleotide that occurs at a specific position in the genome with at least certain frequency (e.g. > 1%).
- **SNV (single-nucleotide variant)**
 - Similar like SNP, but no frequency requirement
- **InDel**
 - Insertion or deletion of one or multiple nucleotides in the genome
- **Structural variation**
 - the variation in structure of the genome, including duplication, insertion, deletion, inversion, copy-number variation, etc.

Identification of SNPs (SNP calling)

Single-nucleotide polymorphism (SNP) is the most common genetic variation among individuals. Next-generation sequencing technology provide a cost-effective tool for SNP detection.

$$P(g|D) = \text{Prior}(g).P(D|g) / \sum \text{Prior}(x).P(D|x)$$

SNP calling algorithm may consider:

- Sequencing quality
- Alignment uniqueness and accuracy
- Likelihood calculation based on observed data
- Prior probability (dbSNP or other resource)

Reference →

Reads

ATGACGGTATGCT
ACGAGAT
ACGAGAT
ACGAGAT
ACGAGAT
ACGGGAT
ACGGGAT
ACGAGAT

Variant calling format (VCF)

```
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=DS,Number=0,Type=Flag,Description="Were any of the samples downsampled?">
##INFO=<ID=Dels,Number=1,Type=Float,Description="Fraction of Reads Containing Spanning Deletions">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=HaplotypeScore,Number=1,Type=Float,Description="Consistency of the site with at most two segregating haplotypes">
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation">
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily the same as the AC), for each ALT allele, in the same order as listed">
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily the same as the AF), for each ALT allele, in the same order as listed">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQ0,Number=1,Type=Integer,Description="Total Mapping Quality Zero Reads">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RPA,Number=.,Type=Integer,Description="Number of times tandem repeat unit is repeated, for each allele (including reference)">
##INFO=<ID=RU,Number=1,Type=String,Description="Tandem repeat unit (bases)">
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias">
##INFO=<ID=STR,Number=0,Type=Flag,Description="Variant is a short tandem repeat">
##contig=<ID=gb|BK006935.2|,length=230218>
##reference=file:///home/lijun/Day1/Variant_calling/S288C_Chromosome_I.fasta
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT exomeSM
gb|BK006935.2| 32 . C T 42.77 . AC=1;AF=0.500;AN=2;BaseQRankSum=0.615;DP=19;Dels=0.00;FS=0.000;HaplotypeScore=13.5644;MLEAC=1;MLEAF=0.500;MQ=60.00;MQ0=0;MQRankSum=-0.280;QD=2.25;ReadPosRankSum=-0.615 GT:AD:DP:GQ:PL 0/1:16,3:19:71:0,504
```

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Mandatory header lines:** ##fileformat=VCFv4.0
- Optional header lines (meta-data about the annotations in the VCF body):** ##fileDate=20100707, ##source=VCFtools, ##reference=NCBI36, ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">, ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">, ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">, ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">, ##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">, ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">, ##ALT=<ID=DEL,Description="Deletion">, ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">, ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
- Reference alleles (GT=0):** A, T, G, C (index 0)
- Alternate alleles (GT>0 is an index to the ALT column):** AT, CT, G, DEL (index 1, 2, 3, 4)
- Phased data (G and C above are on the same chromosome):** 0|1:100
- Deletion:**
- SNP:** C to T
- Large SV:** SVTYPE=DEL;END=300
- Insertion:** A, AT
- Other event:** H2;AA=T

Criteria to filter SNPs (or other variants)

- Hard filter
 - QUAL (Phrep quality) 20
 - DP (Depth) 5
 - QD (QualByDepth) 2
 - AC (Total number of alternate alleles called) 2
 - RPB (Mann-Whitney U test of Read Position Bias (bigger is better))
 - FS (FisherStrand) 60
 - MQ (RMSMappingQuality) 40
 - MQRankSum (MappingQualityRankSumTest) 12.5
 - ReadPosRankSum (ReadPosRankSumTest) 8.0
- Soft filter
 - Variant Quality Score Recalibration (VQSR)
 - Add filter information instead of remove variants

Softwares for SNP calling

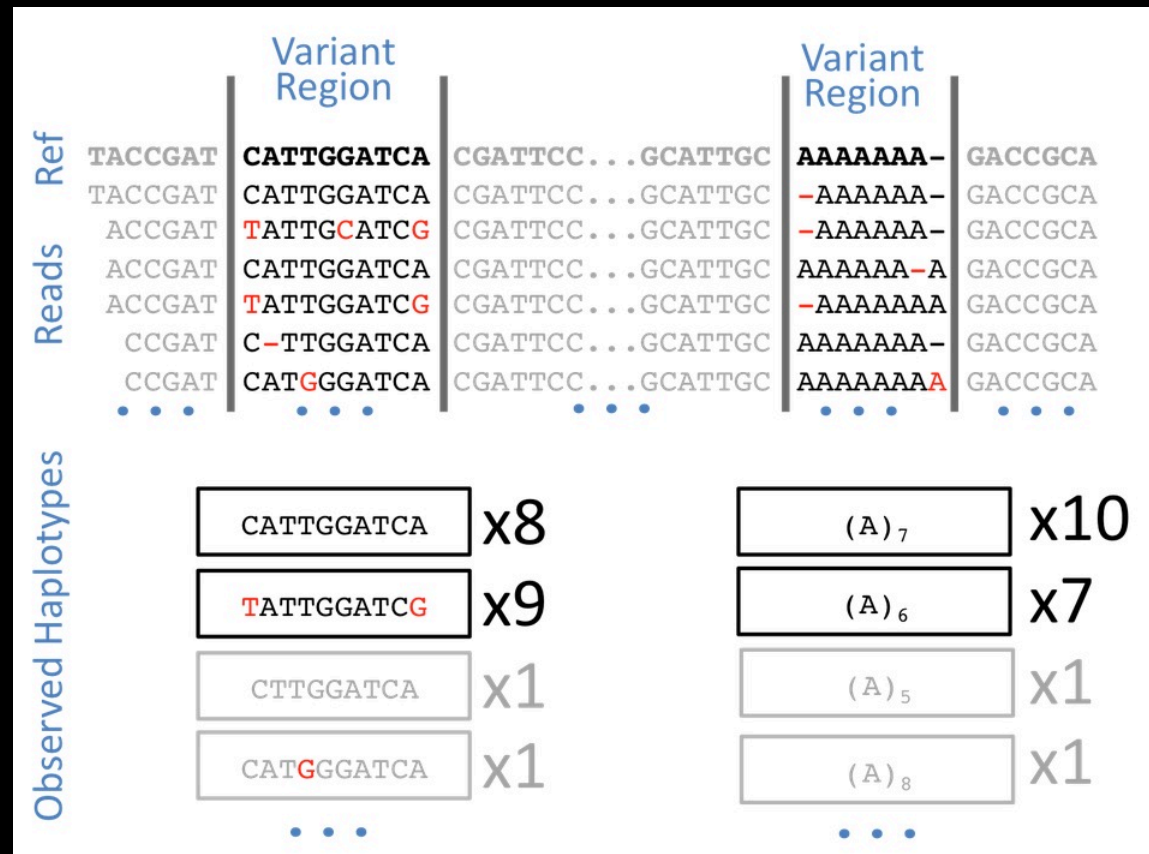
- Samtools/bcftools
 - Complete solution for variant discovery
 - Hard and soft filter
 - Multiplody supported
- GATK
 - Complete solution for variant discovery
 - Both hard filter and soft (self-learned) filter (e.g. human data)
 - Multiplody supported
- Freebayes
 - Variant calling
 - Hard filter
 - Multiplody supported, **population or pool sample supported**

Variant calling in metagenome

- Difficulty
 - a mixture of genome DNA
 - Highly variable abundance
 - Multiple strains for each species
 - Higher uncertainty in alignment
- Software available
 - Freebayes

Why Freebayes works better for calling variants in metagenomes?

- Not precise alignment based
- Flexible ploidy
- Capable to deal with pool samples
- Population diversity incorporated



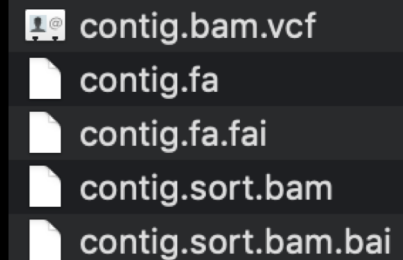
Important parameters in Freebayes

- -T (theta) -- The expected mutation rate or pairwise nucleotide diversity
- -p -- ploidy
- -J -- pooled-discrete
- -K --pooled-continuous

Visualize the variants

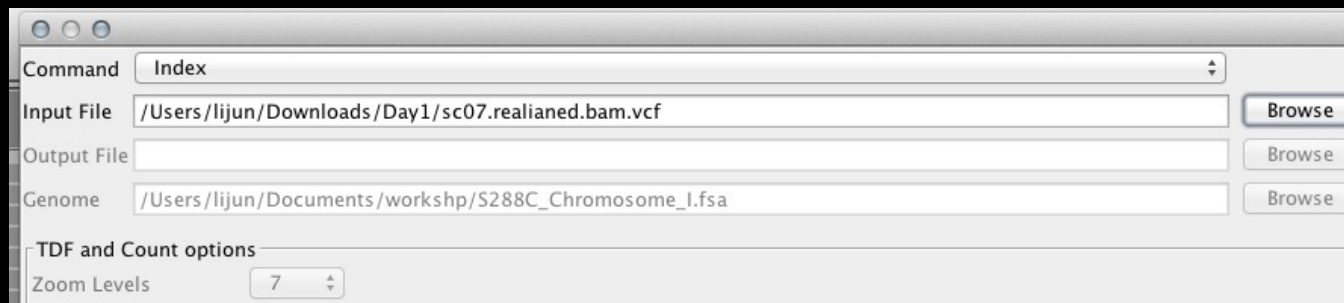
- Open IGV and go to “Genome”-> “load genome from file” , then select “contig.fa” in your pc

Transfer these files to your pc



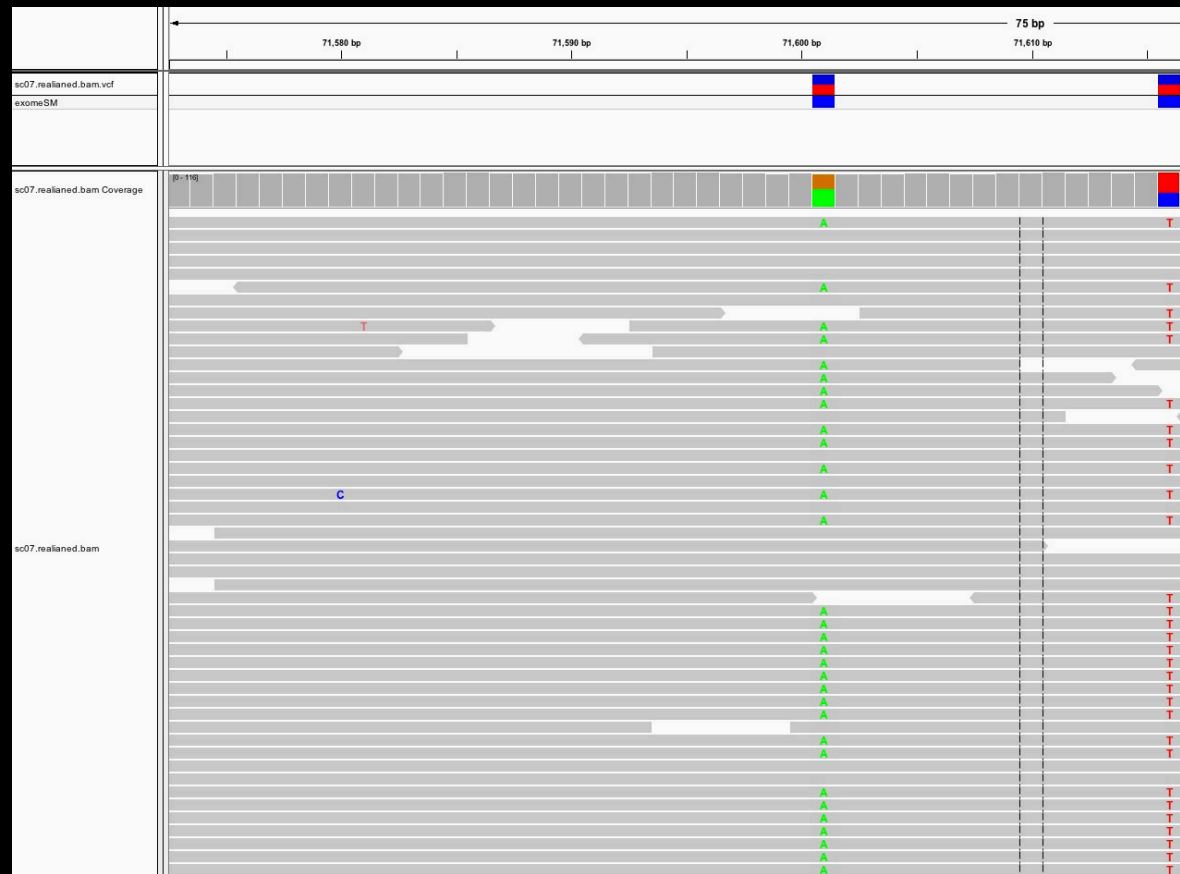
- contig.bam.vcf
- contig.fa
- contig.fa.fai
- contig.sort.bam
- contig.sort.bam.bai

- Load bam alignment “contig.sort.bam”
- Go to “Tools” -> “Run igvtools”, “Browse” your vcf file “contig.bam.vcf” and select command “Index”. Click “Run”



- Load vcf file “contig.bam.vcf”

What you see is like



Practice

- Map yeast reads to the reference genome and refine the alignment
- Call variants using bcftools
- Filter the variants
- Call variants in the demo metagenome data
- Filter the variants using vcffilter
- Visualize the variants