gatk

The current GATK version is 3.4-46

Howdy, Stranger!
It looks like you're new here. If you want to get involved, click one of these buttons!

Sign In Register

Categories
Recent Discussions
Activity
Groups
Participated
Unanswered 189
Best Of...

Categories

All Categories	5.1K
Announcements	139
Ask the GATK team	4K
GATK Documentation Guide	394
FAQs	44
Common Problems	10
Tutorials	21
Presentations	12
Methods and Algorithms	31

Read the Version Highlights for the new GATK version 3.4!

Purpose and operation of Read-backed Phasing

delangel Posts: **71** GATK Dev mod

July 2012 edited May 15 in Methods and Algorithms

This document describes the underlying concepts of physical phasing as applied in the ReadBackedPhasing tool. For a complete, detailed argument reference, refer to the tool documentation page.

Note that as of GATK 3.3, physical phasing is performed automatically by HaplotypeCaller when it is run in `-ERC GVCF` or `-ERC BP_RESOLUTION` mode, so post-processing variant calls with ReadBackedPhasing is no longer necessary unless you want to merge consecutive variants into MNPs.

Underlying concepts

The biological unit of inheritance from each parent in a diploid organism is a set of single chromosomes, so that a diploid organism contains a set of pairs of corresponding chromosomes. The full sequence of each inherited chromosome is also known as a haplotype. It is critical to ascertain which variants are associated with one another in a particular individual. For example, if an individual's DNA possesses two consecutive heterozygous sites in a protein-coding sequence, there are two alternative scenarios of how these variants interact and affect the phenotype of the individual. In one scenario, they are on two different chromosomes, so each one has its own separate effect. On the other hand, if they co-occur on the same chromosome, they are thus expressed in the same protein molecule; moreover, if they are within the same codon, they are highly likely to encode an amino acid that is non-synonymous (relative to the other chromosome). The ReadBackedPhasing program serves to discover these haplotypes based on high-throughput sequencing reads.

How it works

The first step in phasing is to call variants ("genotype calling") using a SAM/BAM file of reads aligned to the reference genome -- this results in a VCF file. Using the VCF file and the SAM/BAM reads file, the ReadBackedPhasing tool considers all reads within a Bayesian framework and attempts to find the local haplotype with the highest probability, based on the reads observed.

The local haplotype and its phasing is encoded in the VCF file as a "|" symbol (which indicates that the alleles of the genotype correspond to the same order as the alleles for the genotype at the preceding variant site). For example, the following VCF indicates that SAMP1 is heterozygous at chromosome 20 positions 332341 and 332503, and the reference base at the first position (A) is on the same chromosome of SAMP1 as the alternate base at the latter position on that chromosome (G), and vice versa (G with C):

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  SAMP1
chr20   332341  rs6076509   A   G   470.60  PASS    AB=0.46;AC=1;AF=0.50;AN=2;DB;DP=52;Dels=0.00;HRun=1;HaplotypeScor
e=0.98;MQ=59.11;MQ0=0;OQ=627.69;QD=12.07;SB=-145.57    GT:DP:GL:GQ 0/1:46:-79.92,-13.87,-84.22:99
chr20   332503  rs6133033   C   G   726.23  PASS    AB=0.57;AC=1;AF=0.50;AN=2;DB;DP=61;Dels=0.00;HRun=1;HaplotypeScor
e=0.95;MQ=60.00;MQ0=0;OQ=894.70;QD=14.67;SB=-472.75    GT:DP:GL:GQ:PQ  1|0:60:-110.83,-18.08,-149.73:99:126.93
```

The per-sample per-genotype PQ field is used to provide a Phred-scaled phasing quality score based on the statistical Bayesian framework employed for phasing. For cases of homozygous sites that lie in between phased heterozygous sites, these homozygous sites will be phased with the same quality as the next heterozygous site.

Note that this tool can only handle diploid data properly. If your organism of interest is polyploid or if you are working with data from pooling experiments, you should not run this tool on your data.

More detailed aspects of semantics of phasing in the VCF format

- The "|" symbol is used for each sample to indicate that each of the alleles of the genotype in question derive from the same haplotype as each of the alleles of the genotype of the same sample in the previous **NON-FILTERED** variant record. That is, rows without FILTER=PASS are essentially ignored in the read-backed phasing (RBP) algorithm.
- Note that the first heterozygous genotype record in a pair of haplotypes will necessarily have a "/" - otherwise, they would be the continuation of the preceding haplotypes.
- A homozygous genotype is always "appended" to the preceding haplotype. For example, any 0/0 or 1/1 record is always converted into 0|0 and 1|1.
- RBP attempts to phase a heterozygous genotype relative the preceding **HETEROZYGOUS** genotype for that sample. If there is sufficient read information to deduce the two haplotypes (for that sample), then the current genotype is declared phased ("/" changed to "|") and assigned a PQ that is proportional to the estimated Phred-scaled error rate. All homozygous genotypes for that sample that lie in between the two heterozygous genotypes are also assigned the same PQ value (and remain phased).
- If RBP cannot phase the heterozygous genotype, then the genotype remains with a "/", and no PQ score is assigned. This site essentially starts a new section of haplotype for this sample.

For example, consider the following records from the VCF file:

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  SAMP1   SAMP2
chr1    1   .   A   G   99  PASS    .   GT:GL:GQ    0/1:-100,0,-100:99  0/1:-100,0,-100:99
chr1    2   .   A   G   99  PASS    .   GT:GL:GQ:PQ 1|1:-100,0,-100:99:60   0|1:-100,0,-100:99:50
chr1    3   .   A   G   99  PASS    .   GT:GL:GQ:PQ 0|1:-100,0,-100:99:60   0|0:-100,0,-100:99:60
chr1    4   .   A   G   99  FAIL    .   GT:GL:GQ    0/1:-100,0,-100:99  0/1:-100,0,-100:99
chr1    5   .   A   G   99  PASS    .   GT:GL:GQ:PQ 0|1:-100,0,-100:99:70   1|0:-100,0,-100:99:60
chr1    6   .   A   G   99  PASS    .   GT:GL:GQ:PQ 0/1:-100,0,-100:99  1|1:-100,0,-100:99:70
chr1    7   .   A   G   99  PASS    .   GT:GL:GQ:PQ 0|1:-100,0,-100:99:80   0|1:-100,0,-100:99:70
chr1    8   .   A   G   99  PASS    .   GT:GL:GQ:PQ 0|1:-100,0,-100:99:90   0|1:-100,0,-100:99:80
```

The proper interpretation of these records is that SAMP1 has the following haplotypes at positions 1-5 of chromosome 1:

```
AGAAA
GGGAG
```

And two haplotypes at positions 6-8:

```
AAA
GGG
```

And, SAMP2 has the two haplotypes at positions 1-8:

```
AAAAGGAA
GGAAAGGG
```

Note that we have excluded the non-PASS SNP call (at chr1:4), thus assuming that both samples are homozygous reference at that site.

Post edited by Geraldine_VdAuwera on May 15

Tagged: readbackedphasing, phasing

Comments

shawpa Posts: **10** Member ★
June 2013

According to what is written above, the readbackphasing doesn't support insertions or deletions. If my vcf file contains indels, should I remove them before running the readbackphasing command. I did run this command without removing indels from my vcf file and I got phasing information from the indel lines in the vcf. Can I not trust these?

Geraldine_VdAuwera Posts: **8,606** Administrator, GATK Dev admin
June 2013