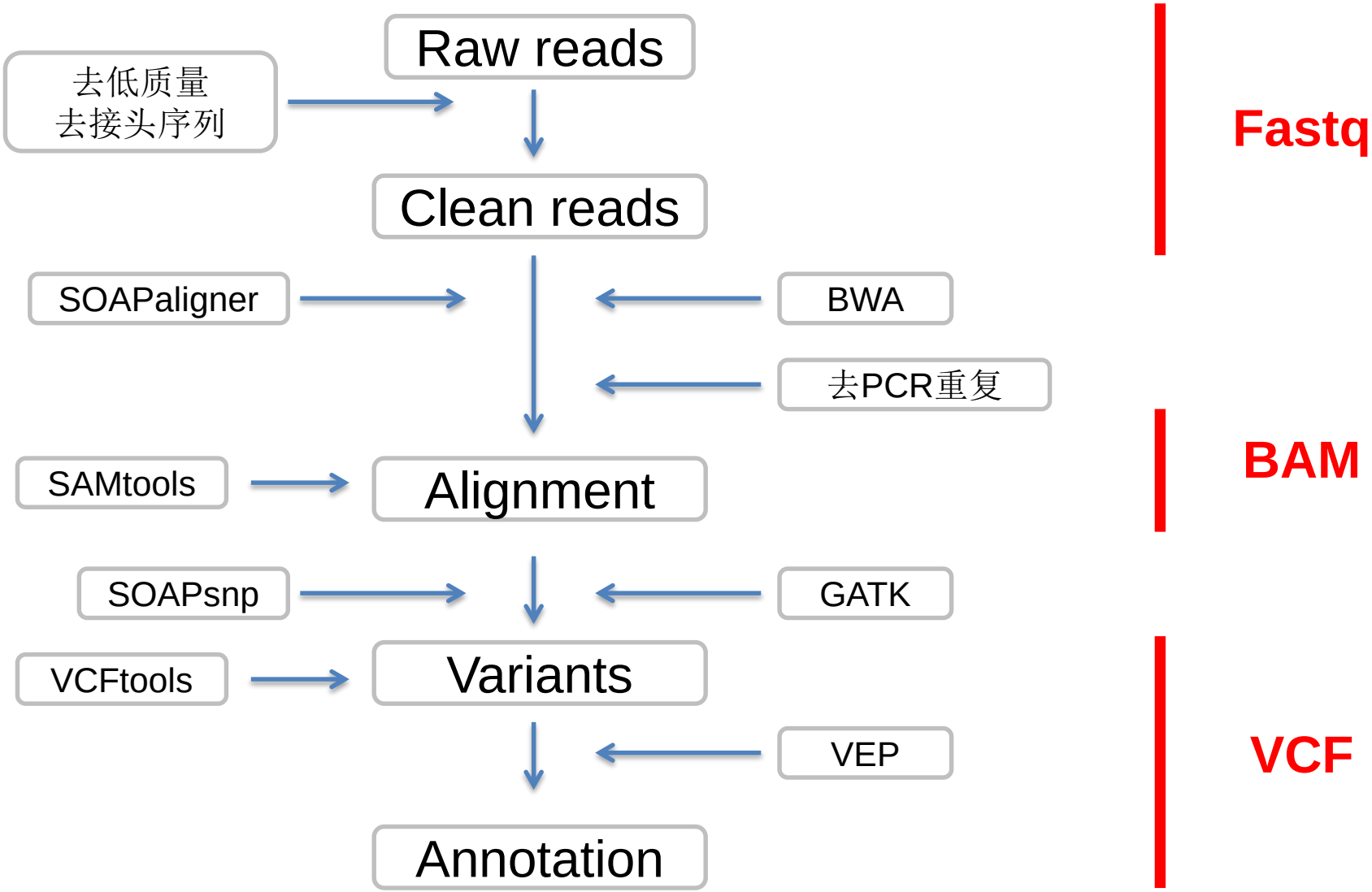
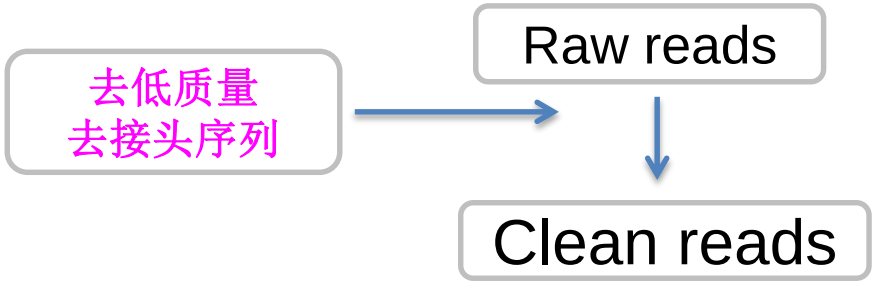
Variant Identification

工 具 篇


Raw reads
去低质量
去接头序列
Clean reads
SOAPaligner
BWA
去PCR重复
SAMtools
Alignment
SOAPsnp
GATK
VCFtools
Variants
VEP
Annotation
Fastq
BAM
VCF




```
soapnuke filter -G -f $adapter1 -r $adapter2 -1 $fastq1 -2 $fastq2 -o $outdir \
-C $clean_read1 -D $read2
```

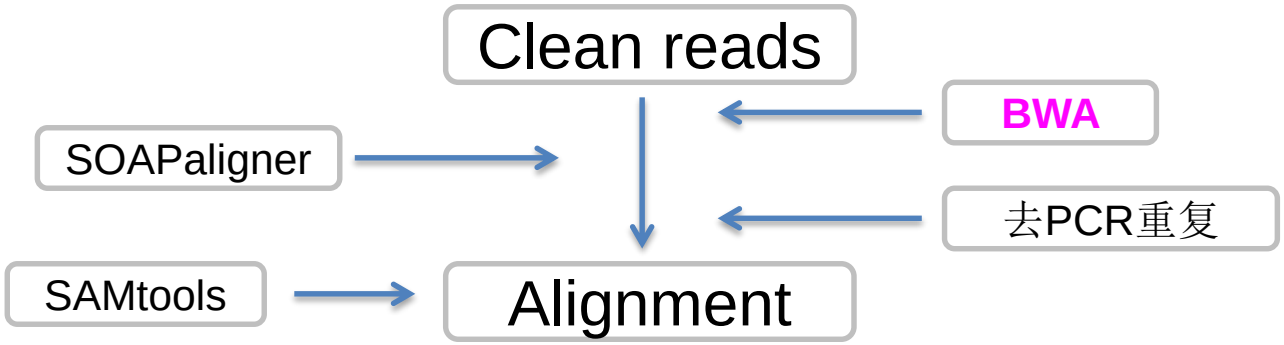
```
[jiangchongyi@login-0-12 ~]$ /ifs2/BC MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Alpha/bin/SOAPnu
ke
                                                                    软件路径
Prpgram: soapnuke
Version: 1.2.0
Contact: YongshengChen<chenyongsheng@genomics.cn>
Command:
         filter        preprocessing sequences   使用的具体工具
         filtersRNA    preprocessing sRNA sequences
         filterDGE     preprocessing DGE sequences
```

涉及参数：

```
-G, --sanger    : <b> set clean data qualtiy system to sanger (default: illumina)
```

为什么要设置成**sanger quality**？

http://sourceforge.net/p/bio-bwa/mailman/message/24412679/




```
[jiangchongyi@login-0-12 ~]$ /ifs2/BC MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Alpha/bin/bwa

Program: bwa (alignment via Burrows-Wheeler transformation)
Version: 0.7.10-r789
Contact: Heng Li <lh3@sanger.ac.uk>

Usage:   bwa <command> [options]

Command: index         index sequences in the FASTA format
         mem           BWA-MEM algorithm
         fastmap       identify super-maximal exact matches
         pemerge       merge overlapping paired ends (EXPERIMENTAL)
         aln           gapped/ungapped alignment
         samse         generate alignment (single ended)
         sampe         generate alignment (paired ended)
         bwasw         BWA-SW for long queries

         fa2pac        convert FASTA to PAC format
         pac2bwt       generate BWT from PAC
         pac2bwtgen    alternative algorithm for generating BWT
         bwtupdate     update .bwt to the new format
         bwt2sa        generate SA from BWT and Occ

Note: To use BWA, you need to first index the genome with `bwa index'.
      There are three alignment algorithms in BWA: `mem', `bwasw', and
      `aln/samse/sampe'. If you are not sure which to use, try `bwa mem'
      first. Please `man ./bwa.1' for the manual.
```

软件路径

新流程中使用的具体工具

BWA三种不同的算法，针对不同的数据类型。

在alignment前需对reference进行index

BWA

http://bio-bwa.sourceforge.net/bwa.shtml

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM.

For 70bp or longer Illumina, 454, Ion Torrent and Sanger reads, assembly contigs and BAC sequences, BWA-MEM is usually the preferred algorithm. For short sequences, BWA-backtrack may be better. BWA-SW may have better sensitivity when alignment gaps are frequent.

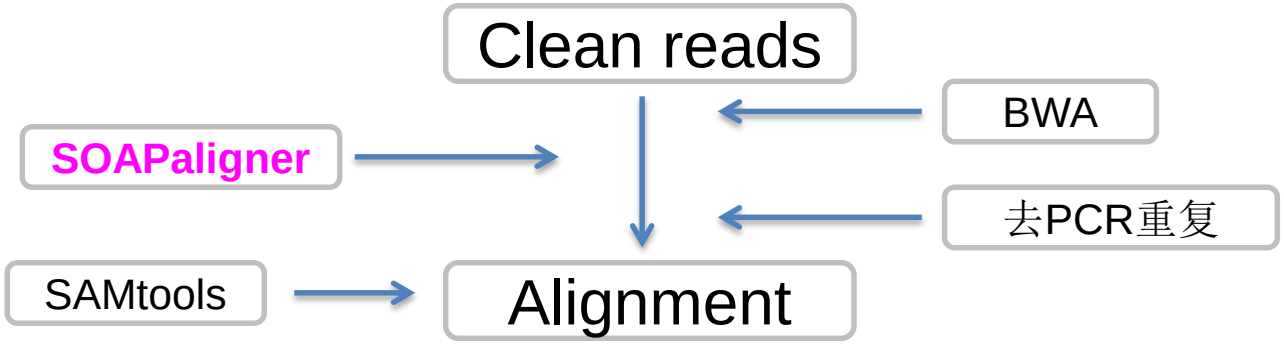
bwa mem -t -R -a -M -C (新流程中的参数)

```
-t INT          number of threads [1]
-R STR          read group header line such as '@RG\tID:foo\tSM:bar' [null]
-a              output all alignments for SE or unpaired PE
-M              mark shorter split hits as secondary
-C              append FASTA/FASTQ comment to SAM output
```

(bwa mem设置insert length)

```
-I FLOAT[,FLOAT[,INT[,INT]]]
              specify the mean, standard deviation (10% of the mean if absent), max
              (4 sigma from the mean if absent) and min of the insert size distribution.
              FR orientation only. [inferred]
```


Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.



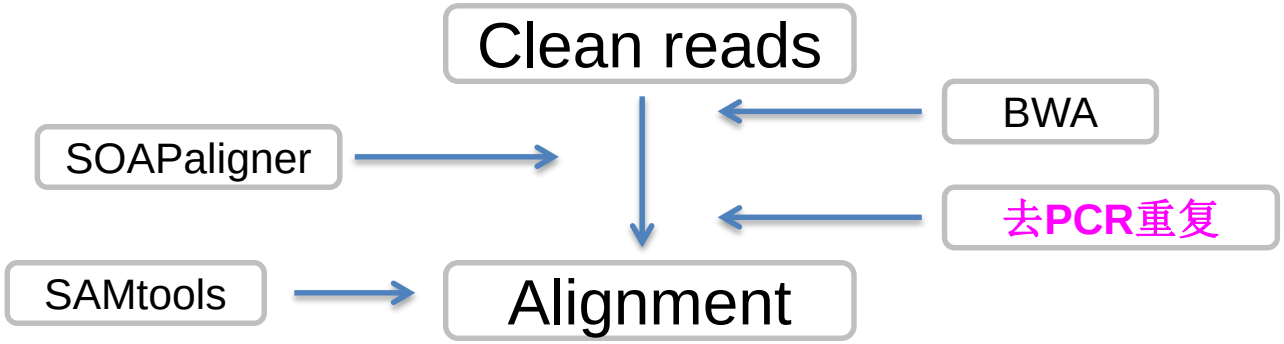

```
[jiangchongyi@compute-23-16 new2]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Alpha/bin/
soap
                                                    软件路径

Program: SOAPaligner/soap2
Compile Date: Sun Aug 22 11:51:04 CST 2010
Author:  BGI shenzhen
Version: 2.21
Contact: soap@genomics.org.cn

Usage:  soap [options]
        -a  <str>   query a file, *.fq, *.fa
        -b  <str>   query b file
        -D  <str>   reference sequences indexing table, *.index format
        -o  <str>   output alignment file(txt)
        -M  <int>   match mode for each read or the seed part of read, which shouldn't contain m
ore than 2 mismaches, [4]
                    0: exact match only
                    1: 1 mismatch match only
                    2: 2 mismatch match only
                    4: find the best hits
        -u  <str>   output unmapped reads file
        -t          output reads id instead reads name, [none]
        -l  <int>   align the initial n bps as a seed [256] means whole length of read
        -n  <int>   filter low-quality reads containing >n Ns before alignment, [5]
        -r  [0,1,2] how to report repeat hits, 0=none; 1=random one; 2=all, [1]
        -m  <int>   minimal insert size allowed, [400]
        -x  <int>   maximal insert size allowed, [600]
        -2  <str>   output file of unpaired alignment hits
        -v  <int>   maximum number of mismatches allowed on a read. [5] bp
        -s  <int>   minimal alignment length (for soft clip) [255] bp
        -g  <int>   one continuous gap size allowed on a read. [0] bp
        -R          for long insert size of pair end reads RF. [none](means FR pair)
        -e  <int>   will not allow gap exist inside n-bp edge of a read, default=5
        -p  <int>   number of processors to use, [1]
```

SOAPaligner

http://soap.genomics.org.cn/soapaligner.html

SOAPaligner/soap2 is a member of the SOAP (Short Oligonucleotide Analysis Package). It is an updated version of SOAP software for short oligonucleotide alignment. The new program features in super fast and accurate alignment for huge amounts of short reads generated by Illumina/Solexa Genome Analyzer. It require only 2 minutes aligning one million single-end reads onto the human reference genome. Another remarkable improvement of SOAPaligner is that it now supports a wide range of the read length. SOAPaligner benefitted in time and space efficiency by a revolution in the basic data structures and algorithms **(2way-BWT)** used.

Li R.Q. et al.(2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics,25:1966-67.



picard MarkDuplicates.jar

```
[jiangchongyi@compute-23-16 picard-tools-1.117]$ java -jar /ifs2/BC_MD/DEV/WorkFlow/Sofware/picard-tools-1.117/MarkDuplicates.jar --help
USAGE: MarkDuplicates [options]

Documentation: http://picard.sourceforge.net/command-line-overview.shtml#MarkDuplicates

Examines aligned records in the supplied SAM or BAM file to locate duplicate molecules. All records are then written to the output file with the duplicate records flagged.
Version: 1.117(107391d3f3e72b31589868c250262ca79659f577_1405353489)
```

软件路径

功能

SAMtools rmdup

```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/bin/samtools rmdup

Usage:   samtools rmdup [-sS] <input.srt.bam> <output.bam>

Option: -s    rmdup for SE reads
        -S    treat PE reads as SE in rmdup (force -s)
```

软件路径

Q: What is the difference between MarkDuplicates and SAMtools rmdup?

A: SAMtools rmdup does not remove interchromosomal duplicates. MarkDuplicates does remove these duplicates.

Picard

http://broadinstitute.github.io/picard/command-line-overview.html

Picard is a set of tools (in Java) for working with next generation sequencing data in the BAM format.

较常用的工具：

AddOrReplaceReadGroups.jar

Replaces all read groups in the INPUT file with a single new read group and assigns all reads to this read group in the OUTPUT BAM.

```
java -jar AddOrReplaceReadGroups.jar I=sample.bam O=sample_addGroup.bam
SORT_ORDER=coordinate CREATE_INDEX=true RGPL=illumina RGID=184
RGSM=sample184 RGLB=bar RGPU=pu184 VALIDATION_STRINGENCY=LENIENT
```

MarkDuplicates.jar

Examines aligned records in the supplied SAM or BAM file to locate duplicate molecules. All records are then written to the output file with the duplicate records flagged.

```
java -Xmx4g -jar MarkDuplicates.jar INPUT=Sample.sorted.bam OUTPUT=Sample.dedup.bam
METRICS_FILE=Sample.dedup.txt VALIDATION_STRINGENCY=LENIENT REMOVE_DUPLICATES=false
```

Picard

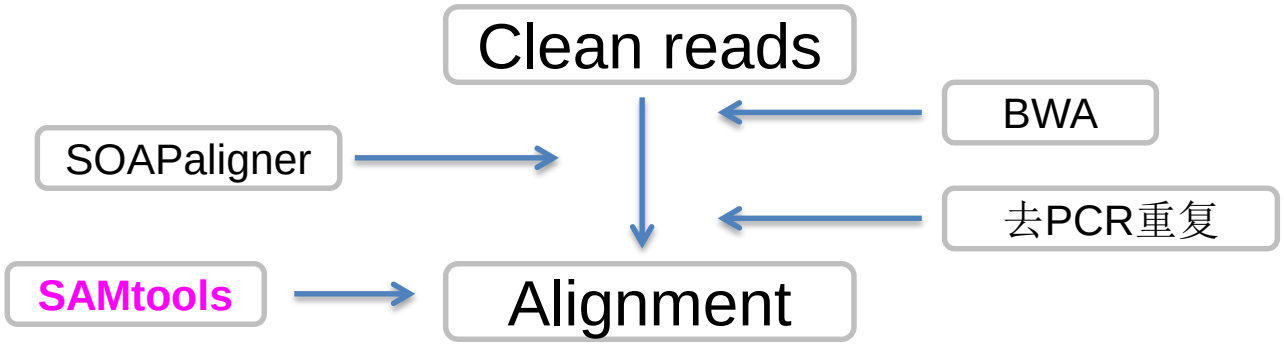
http://sourceforge.net/p/picard/wiki/Main_Page/

Q: How does MarkDuplicates work?

A: Essentially what it does (for pairs; single-end data is also handled) is to find the 5' coordinates and mapping orientations of each read pair. When doing this it takes into account all clipping that has taking place as well as any gaps or jumps in the alignment. You can thus think of it as determining "if all the bases from the read were aligned, where would the 5' most base have been aligned". It then matches all read pairs that have identical 5' coordinates and orientations and marks as duplicates all but the "best" pair. "Best" is defined as the read pair having the highest sum of base qualities as bases with Q >= 15.

Q: A Picard program complains that CIGAR M operator maps off the end of reference. I want this record to be treated as valid despite the fact that the alignment end is greater than the length of the reference sequence.

A: Picard validation errors may be turned into warnings by passing the command line argument VALIDATION_STRINGENCY=LENIENT. Picard validation messages may be suppressed completely with VALIDATION_STRINGENCY=SILENT. Another option is to use CleanSam to soft-clip these reads so they don't map off the end of the reference.




```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-44428cd

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         mpileup     multi-way pileup
         depth       compute the depth
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         idxstats    BAM index stats (r595 or later)
         fixmate     fix mate information
         flagstat    simple stats
         calmd       recalculate MD/NM tags and '=' bases
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
         reheader    replace BAM header
         cat         concatenate BAMs
         bedcov      read depth per BED region
         targetcut   cut fosmid regions (for fosmid pool only)
         phase       phase heterozygotes
         bamshuf     shuffle and group alignments by name
```

软件路径

功能

较常用工具

SAMtools

http://www.htslib.org/doc/samtools-1.2.html

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

较常用的工具：

```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools view

Usage:   samtools view [options] <in.bam>|<in.sam> [region1 [...]]

Options: -b       output BAM
         -h       print header for the SAM output
         -H       print header only (no alignments)
         -S       input is SAM
         -u       uncompressed BAM output (force -b)
         -1       fast compression (force -b)
         -x       output FLAG in HEX (samtools-C specific)
         -X       output FLAG in string (samtools-C specific)
         -c       print only the count of matching records
         -B       collapse the backward CIGAR operation
         -@ INT   number of BAM compression threads [0]
         -L FILE  output alignments overlapping the input BED FILE [null]
         -t FILE  list of reference names and lengths (force -S) [null]
         -T FILE  reference sequence file (force -S) [null]
         -o FILE  output file name [stdout]
         -R FILE  list of read groups to be outputted [null]
         -f INT   required flag, 0 for unset [0]
         -F INT   filtering flag, 0 for unset [0]
         -q INT   minimum mapping quality [0]
         -l STR   only output reads in library STR [null]
         -r STR   only output reads in read group STR [null]
         -s FLOAT fraction of templates to subsample; integer part as seed [-1]
         -?       longer help
```

samtools view sample.bam | less -S

1.使用管道

2.该命令只显示alignment的情况，如需header信息，用samtools view -h sample.bam | less 这样既有header也有alignment信息；

有时会用到-q输出高质量的alignment信息

通过-f或-F设置输出特定的alignment信息，详见SAM格式介绍；

SAMtools

http://www.htslib.org/doc/samtools-1.2.html

较常用的工具：

```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools sort

Usage:   samtools sort [options] <in.bam> <out.prefix>

Options: -n         sort by read name
         -f         use <out.prefix> as full file name instead of prefix
         -o         final output to stdout
         -l INT     compression level, from 0 to 9 [-1]
         -@ INT     number of sorting and compression threads [1]
         -m INT     max memory per thread; suffix K/M/G recognized [768M]

[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools index
Usage: samtools index <in.bam> [out.index]
```

处理bam文件的下游工具一般都要求先对进行sort和index

samtools sort sample.bam sample.sorted

samtools index sample.sorted.bam

```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools merge

Usage:   samtools merge [-nr] [-h inh.sam] <out.bam> <in1.bam> <in2.bam> [...]

Options: -n         sort by read names
         -r         attach RG tag (inferred from file names)
         -u         uncompressed BAM output
         -f         overwrite the output BAM if exist
         -1         compress level 1
         -l INT     compression level, from 0 to 9 [-1]
         -@ INT     number of BAM compression threads [0]
         -R STR     merge file in the specified region STR [all]
         -h FILE    copy the header in FILE to <out.bam> [in1.bam]

Note: Samtools' merge does not reconstruct the @RG dictionary in the header. Users
      must provide the correct header with -h, or uses Picard which properly maintains
      the header dictionary in merging.
```

samtools的merge通常用在把同一个sample不同染色体的bam文件合在一起；

注意输出文件在前面，后面才跟着输入文件！

merge后得到bam文件的header默认为in1.bam的header；如果来源不同的bam合在一起时需提供合适的header文件，否则下游工具用到header信息(@RG)时会报错；

SAMtools

http://www.htslib.org/doc/samtools-1.2.html

较常用的工具：

```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools mpileup

Usage: samtools mpileup [options] in1.bam [in2.bam [...]]
```

1. Generate VCF, BCF or pileup for one or multiple BAM files. Alignment records are grouped by sample (SM) identifiers in @RG header lines. If sample identifiers are absent, each input file is regarded as one sample.
2. Pileup format (without **-u** or **-g**): **http://samtools.sourceforge.net/pileup.shtml**
3. 新流程使用samtools mpileup进行INDELs calling；

```
-B              disable BAQ computation
```

[Samtools-announce] New feature: Base Alignment Quality (BAQ)

From: Heng Li <lh3@sa...> - 2010-10-13 14:08:09

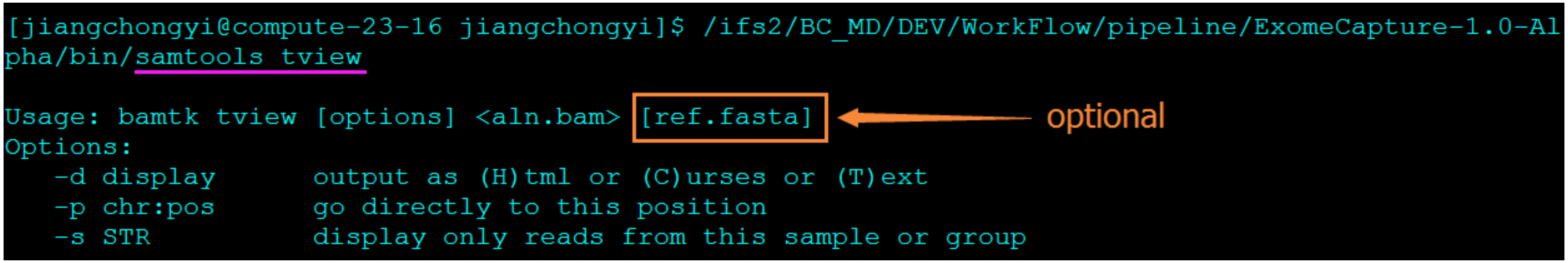
I seldom send email to samtools-announce, but I think it is worth it this time.

Samtools now calculates a per-base alignment quality (BAQ) which directly measures the probability of a read base (not the entire read, so different from mapping quality) being wrongly aligned. It is designed to resolve false SNPs caused by misalignment, especially around indels and in low-complexity regions. Application to data from the 1000 genomes project has confirmed its effectiveness. The theory is in http://lh3lh3.users.sourceforge.net/download/samtools.pdf.

SAMtools

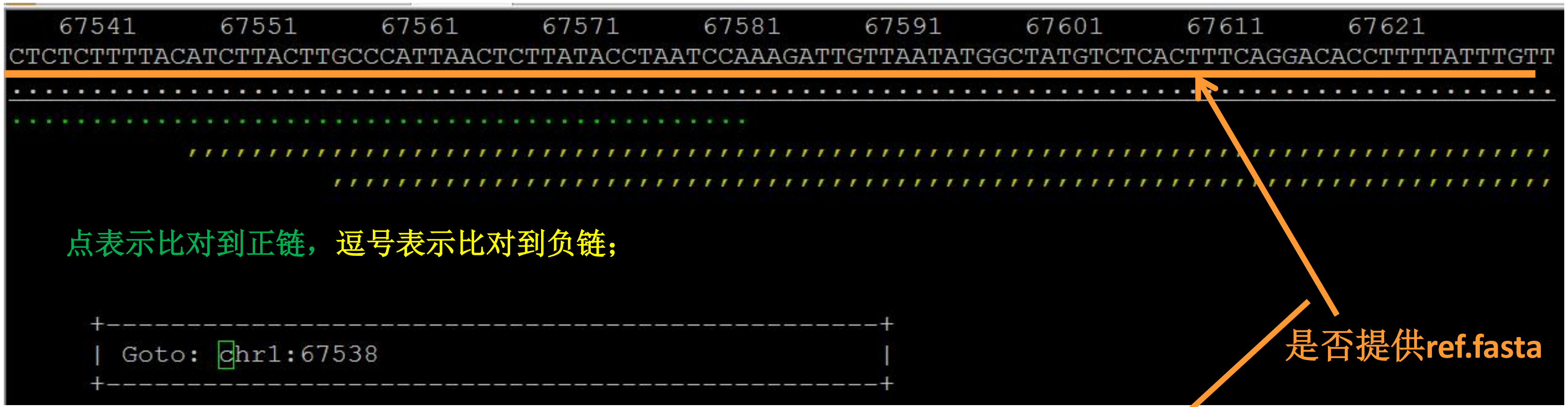
http://www.htslib.org/doc/samtools-1.2.html

较常用的工具：



```
[jiangchongyi@compute-23-16 jiangchongyi]$ /ifs2/BC_MD/DEV/WorkFlow/pipeline/ExomeCapture-1.0-Al
pha/bin/samtools tview

Usage: bamtk tview [options] <aln.bam> [ref.fasta]
Options:
   -d display      output as (H)tml or (C)urses or (T)ext
   -p chr:pos      go directly to this position
   -s STR          display only reads from this sample or group
```

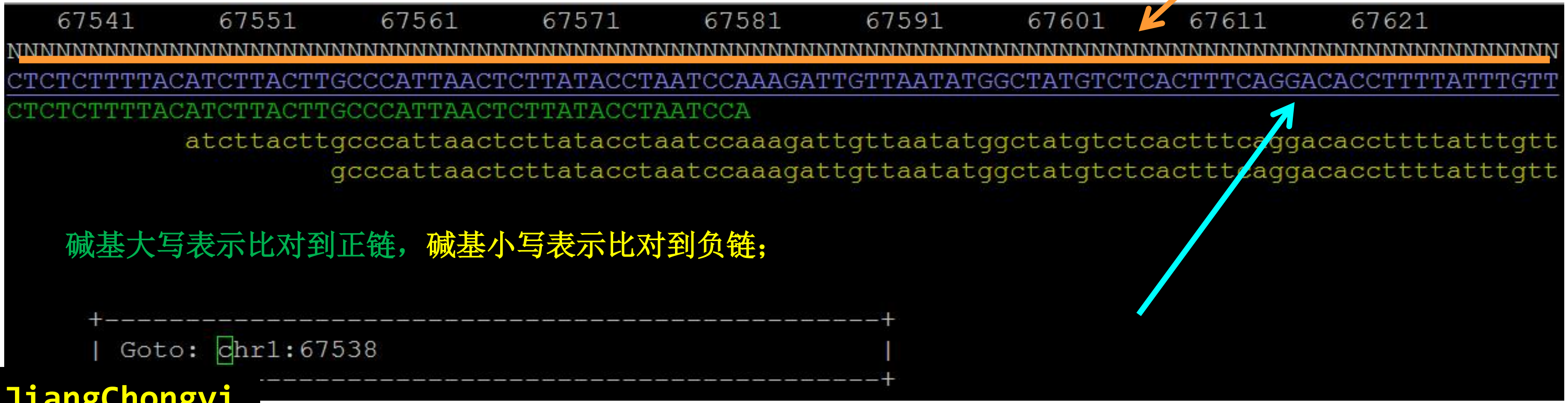
optional



```
   67541     67551     67561     67571     67581     67591     67601     67611     67621
CTCTCTTTTACATCTTACTTGCCCATTAACTCTTATACCTAATCCAAAGATTGTTAATATGGCTATGTCTCACTTTCAGGACACCTTTTATTTGTT
................................................................................................
..............................................
           ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                    ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,



     +----------------------------------------------+
     | Goto: chr1:67538                             |
     +----------------------------------------------+
```

点表示比对到正链，逗号表示比对到负链；

是否提供ref.fasta



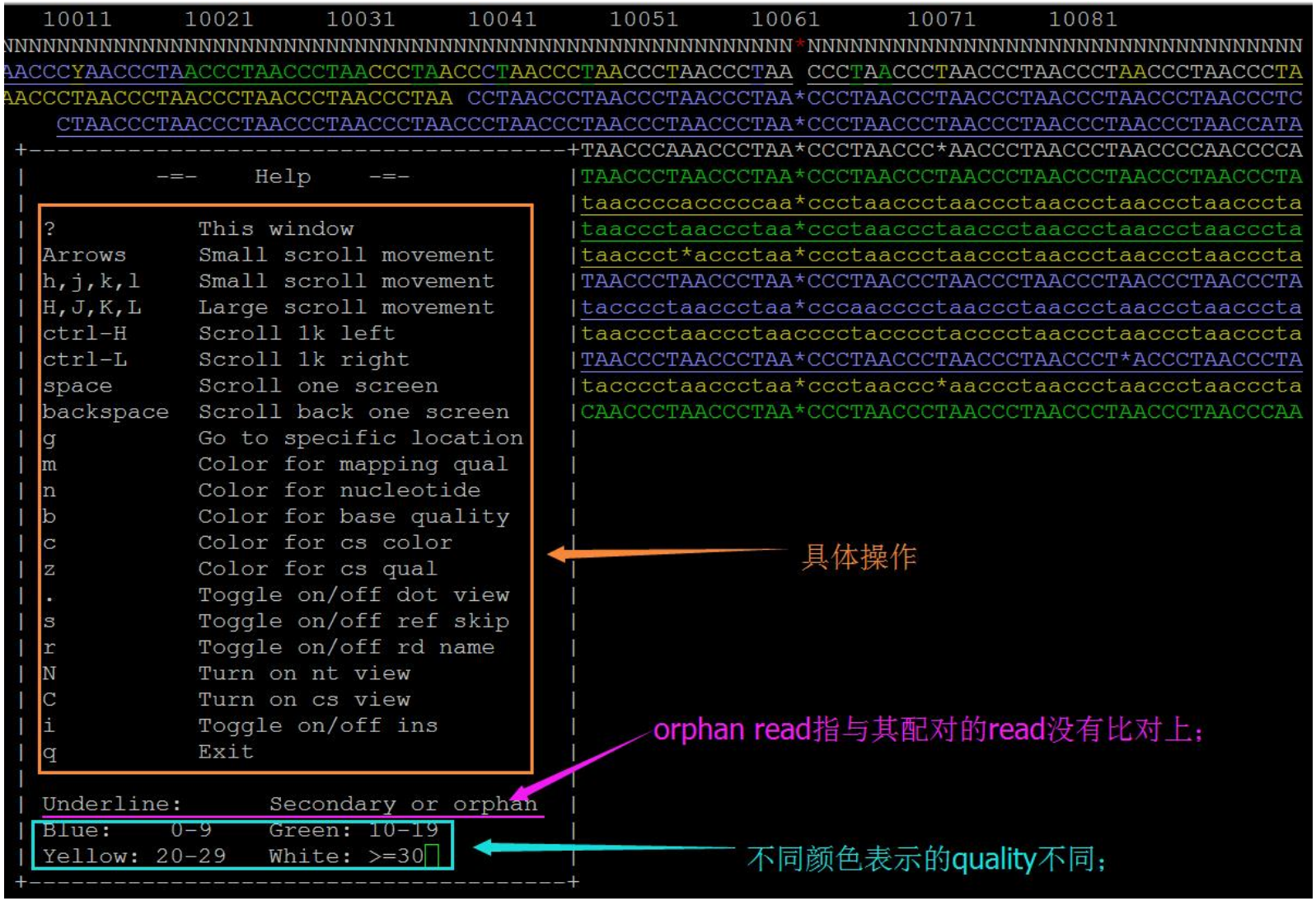
```
   67541     67551     67561     67571     67581     67591     67601     67611     67621
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CTCTCTTTTACATCTTACTTGCCCATTAACTCTTATACCTAATCCAAAGATTGTTAATATGGCTATGTCTCACTTTCAGGACACCTTTTATTTGTT
CTCTCTTTTACATCTTACTTGCCCATTAACTCTTATACCTAATCCA
           atcttacttgcccattaactcttatacctaatccaaagattgttaatatggctatgtctcactttcaggacaccttttatttgtt
                    gcccattaactcttatacctaatccaaagattgttaatatggctatgtctcactttcaggacaccttttatttgtt



     +----------------------------------------------+
     | Goto: chr1:67538                             |
     +----------------------------------------------+
```

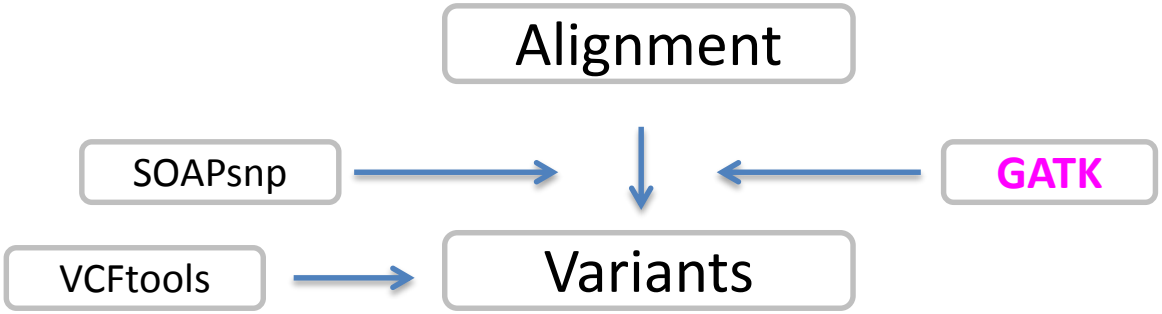
碱基大写表示比对到正链，碱基小写表示比对到负链；

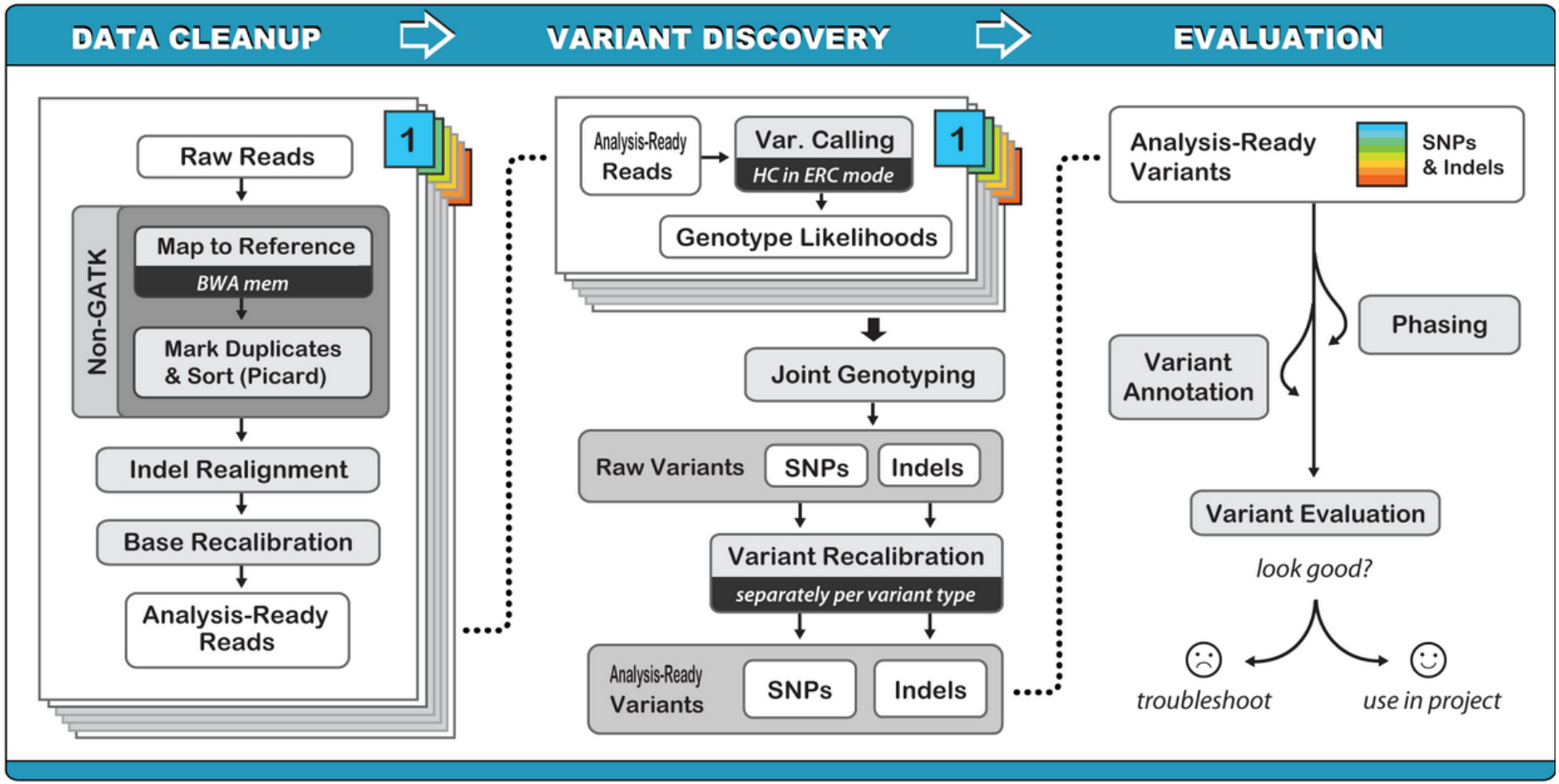
SAMtools

http://www.htslib.org/doc/samtools-1.2.html

较常用的工具：



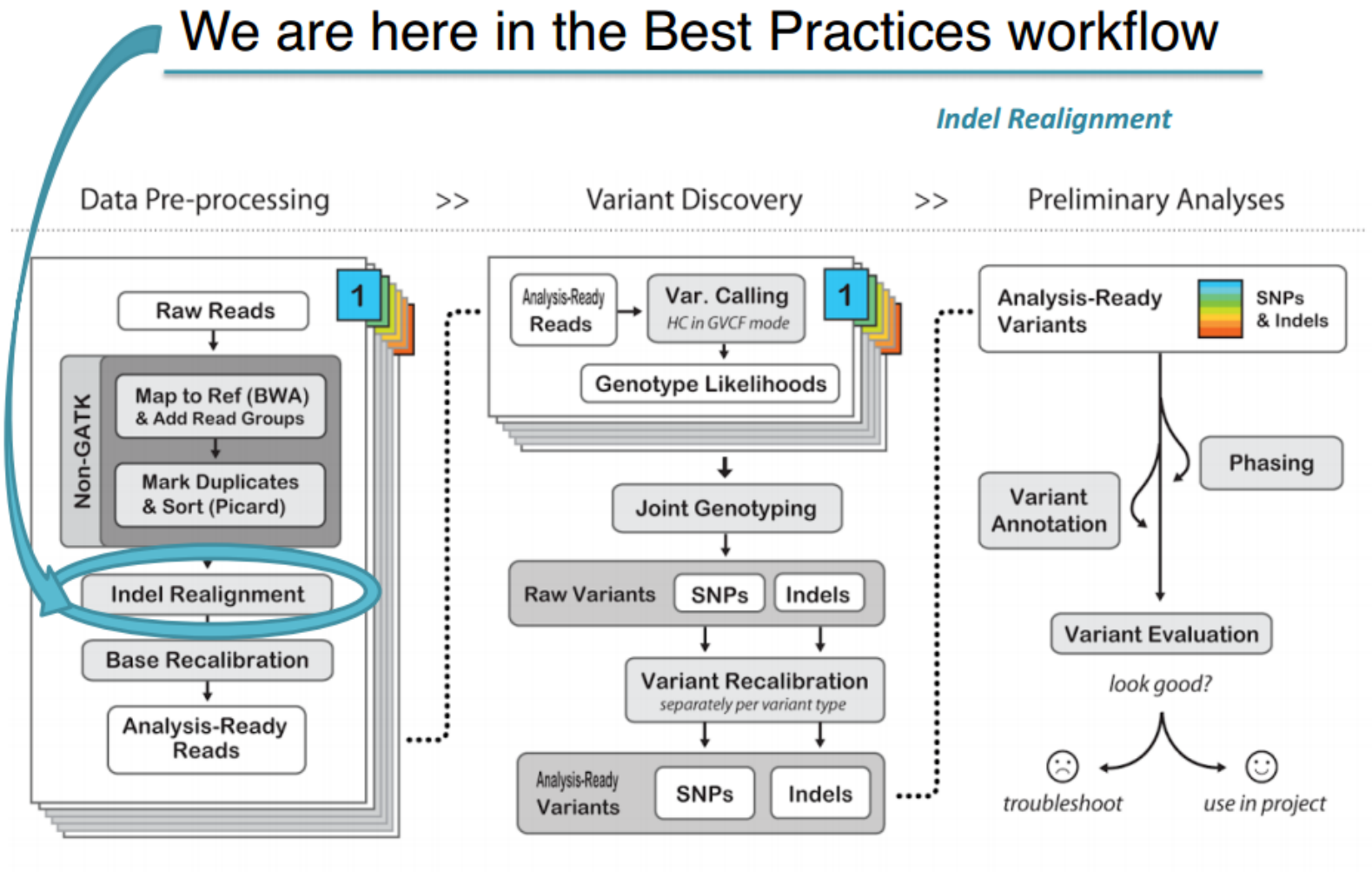




Emma_JiangChongyi

GATK Indel Realignment

https://www.broadinstitute.org/gatk/



gatk

Emma_JiangChongyi

GATK Indel Realignment

https://www.broadinstitute.org/gatk/

Purpose: Improving the original alignments of the reads based on multiple sequence (re-)alignment.

Why this step?

a. InDels in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches;
b. Since read mapping algorithms operate on each read independently, it is impossible to place reads on the reference genome such at mismatches are minimized across all reads. Consequently, even when some reads are correctly mapped with indels, reads covering the indel near just the start or end of the read are often incorrectly mapped with respect the true indel, also requiring realignment;
c. These artifactual mismatches can harm base quality recalibration and variant detection (unless a sophisticated caller like the Haplotype Caller is used).

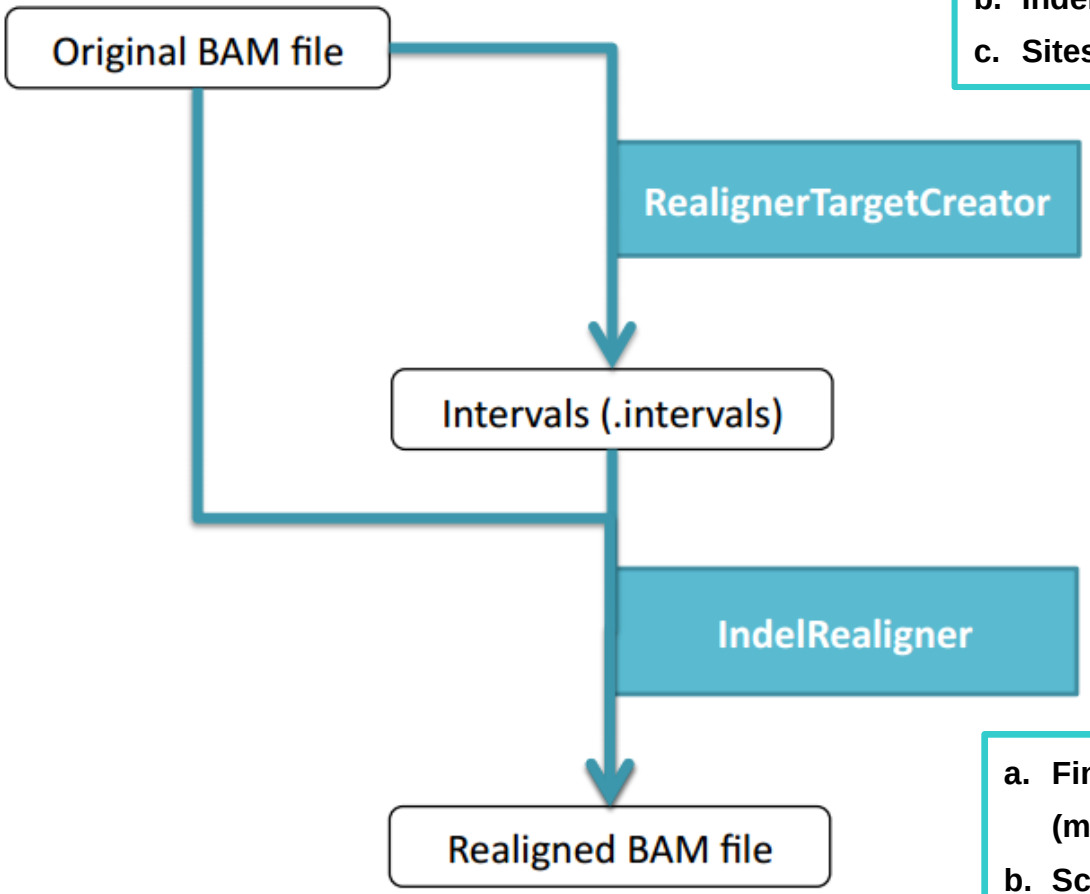
Indel Realignment steps/tools

a. Identify what regions need to be realigned (RealignerTargetCreator)
b. Perform the actual realignment (IndelRealigner)

GATK Indel Realignment

https://www.broadinstitute.org/gatk/

Workflow:



a. Known sites (e.g. dbSNP, 1000 Genomes)
b. Indels seen in original alignment (in CIGAEs)
c. Sites where evidence suggests a hidden indel

a. Find the best alternate consensus sequence (maximum of 1 indel);
b. Score of the alternate consensus is the total sum of the qualtiy scores of mismatching bases;
c. If the score of the best alternate consensus is sufficiently better than the original alignment, then the proposed realignment is accepted;

GATK

Q: What should I use as known variants/sites for running tool X?
A: Each tool uses known sites differently, but what is common to all is that they use them to help distinguish true variants from false positives, which is very important to how these tools work. If you don't provide known sites, the statistical analysis of the data will be skewed, which can dramatically affect the sensitivity and reliability of the results.

	dbSNP	Mills_and_1000G_gold_standard.indel.b37.sites.vcf	1000G_phase1.indels.b37.vcf	HapMap	Omni
RealignerTargetCreator		√	√		
IndelRealigner		√	√		
BaseRecalibrator	√	√	√		
UC/HC	√				
VariantRecalibrator	√	√	√	√	√
VariantEval	√				

https://www.broadinstitute.org/gatk/guide/article?id=1247

GATK Indel Realignment

https://www.broadinstitute.org/gatk/

Evaluation:

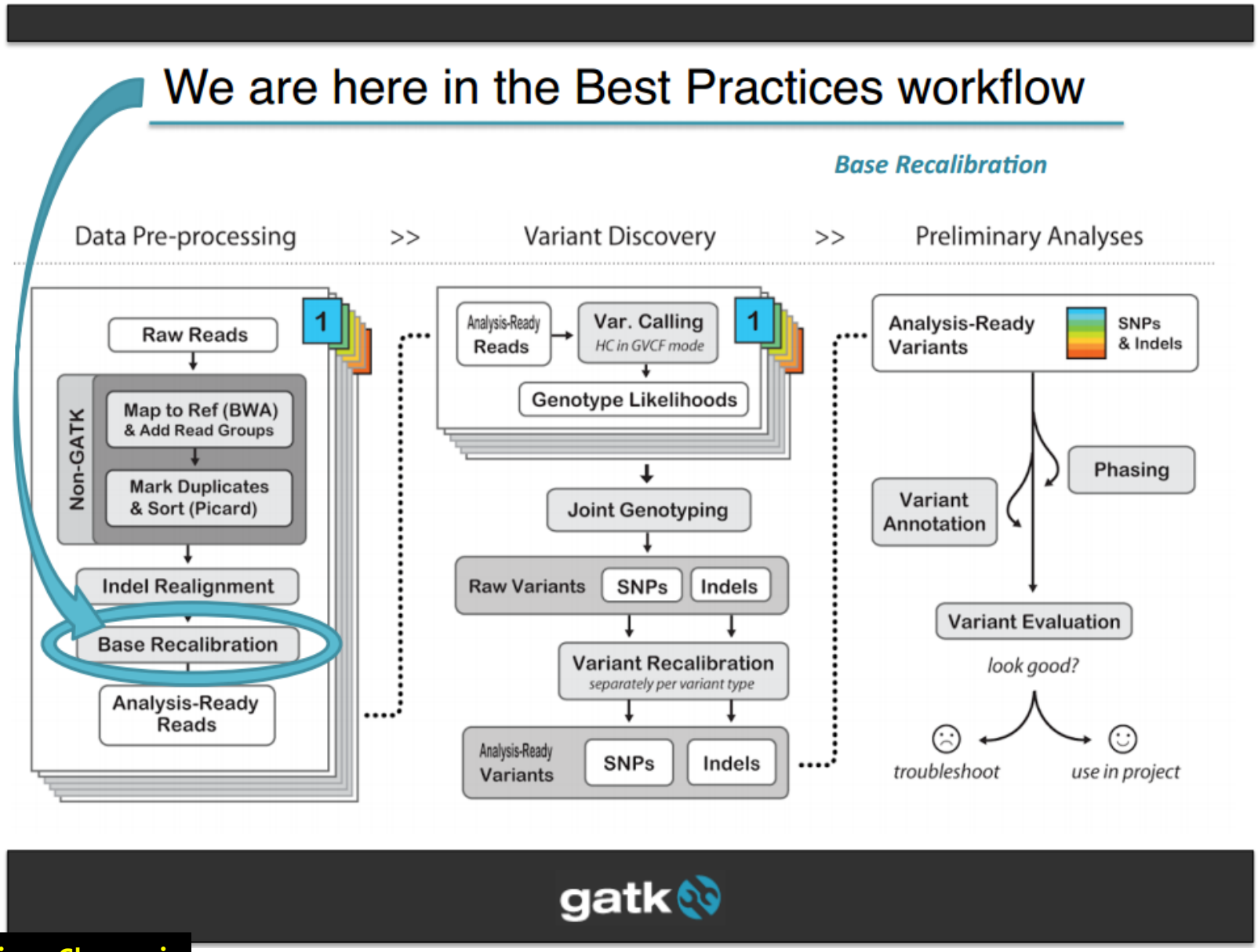
a. Indel realigner changes the CIGAR string of reaglinged reads but maintains the original CIGAR (with OC tag);
b. But no formal measure to assess the accuracy or completeness of the realignment process;
c. Latest tools being implemented for discovering mutations all include some sort of assembly step (for which upstream realignment is not really helpful);
d. But big improvement for Base Quality Score Recalibration when run on realigned BAM files;
e. Also still useful for legacy tools, e.g. full realignment should be performed if using the GATK's Unified Genotyper for calling variants;

Command lines:

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R ref.fasta \
-I input.bam -o forIndelRealigner.intervals [--known /path/to/indels.vcf]
java -Xmx4g -jar GenomeAnalysisTK.jar -T IndelRealigner -R ref.fasta -I input.bam \
-targetIntervals intervalListFromRTC.intervals -o realignedBam.bam \
[-known /path/to/indels.vcf]
```

GATK Base Quality Score Recalibration

https://www.broadinstitute.org/gatk/



Emma_JiangChongyi

GATK Base Quality Score Recalibration

https://www.broadinstitute.org/gatk/

Purpose: Assigning accurate confidence scores to each sequenced base.

Why this step?

a. Quality scores issued by sequencers are inaccurate and biased;
b. Quality scores are critical for all downstream analysis;
c. Systematic biases are a major contributor to bad calls;
d. BQSR identifies patterns in how errors correlate with base features;

How BQSR works?

a. Empowered by looking at entire lane of data;
b. Analyze covariation among several features of a base, e.g.:
   - Reported qualtiy score;
   - Position within the read (machine cycle);
   - Preceding and current nucleotide (sequencing chemistry effect);
c. Apply corrections to recalibrate the quality scores of all reads in the BAM files based on the patterns identified;

GATK Base Quality Score Recalibration

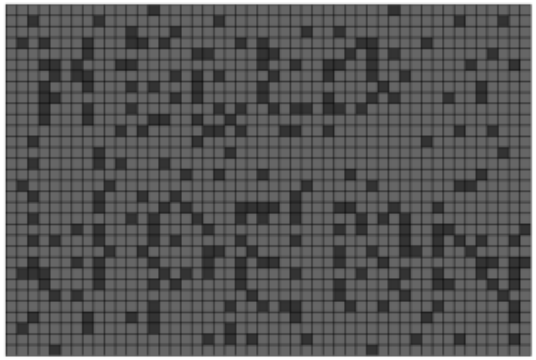
https://www.broadinstitute.org/gatk/

How covariates are analyzed to identify patterns?

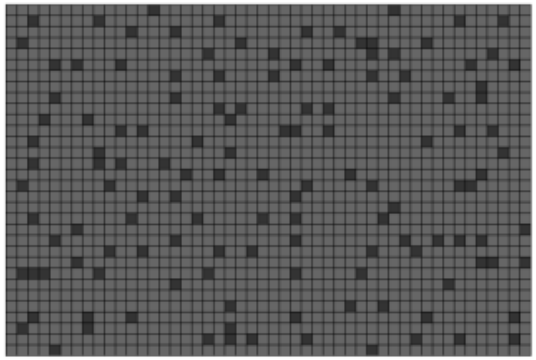
a. Any sequence mismatch = **error** ***except known variants!***

b. Keep track of number of observations and number of **errors** as a function of various error covariates (lane, original quality score, machine cycle, and sequencing context;

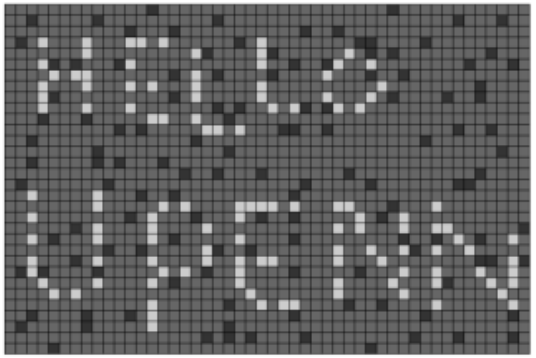
Mask out most of the true variation



All mismatches



Known variation



GATK Base Quality Score Recalibration

https://www.broadinstitute.org/gatk/

Apply corrections to recalibrate the quality scores

For each base in each read:

-- is it in AA context? -> adjust by X points

-- ...

-- is it at 3rd position? -> adjust by Y points

-- ...

```
#:GATKTable:6:3:%s:%s:%.4f:%.4f:%d:%.2f:;
#:GATKTable:RecalTable0:
ReadGroup        EventType  EmpiricalQuality  EstimatedQReported
exampleBAM.bam   M                   17.0000             17.0000
exampleBAM.bam   I                   45.0000             45.0000
exampleBAM.bam   D                   45.0000             45.0000

#:GATKTable:6:3:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable1:
ReadGroup        QualityScore  EventType  EmpiricalQuality  Observ
exampleBAM.bam             17  M                   17.0000
exampleBAM.bam             45  I                   45.0000
exampleBAM.bam             45  D                   45.0000

#:GATKTable:8:556:%s:%s:%s:%s:%s:%.4f:%d:%.2f:;
#:GATKTable:RecalTable2:
ReadGroup        QualityScore  CovariateValue  CovariateName  EventType  EmpiricalQuality  Observations  Errors
exampleBAM.bam             17  AA              Context        M                   17.0000            18    0.00
exampleBAM.bam             17  CA              Context        M                   17.0000            23    0.00
exampleBAM.bam             17  GA              Context        M                   17.0000            18    0.00
exampleBAM.bam             17  TA              Context        M                   17.0000            22    2.00
exampleBAM.bam             17  AC              Context        M                   17.0000             9    0.00
exampleBAM.bam             17  CC              Context        M                   17.0000            13    0.00
exampleBAM.bam             17  GC              Context        M                   17.0000            13    2.00
exampleBAM.bam             17  TC              Context        M                   17.0000            22    2.00
exampleBAM.bam             17  AG              Context        M                   17.0000            23    0.00
exampleBAM.bam             17  CG              Context        M                   17.0000             5    0.00
exampleBAM.bam             17  GG              Context        M                   17.0000            42    0.00
exampleBAM.bam             17  TG              Context        M                   17.0000            35    3.00
exampleBAM.bam             17  AT              Context        M                   17.0000            30    0.00
exampleBAM.bam             17  CT              Context        M                   17.0000            19    0.00
exampleBAM.bam             17  GT              Context        M                   17.0000            26    0.00
exampleBAM.bam             17  TT              Context        M                   17.0000            45    2.00
exampleBAM.bam             45  AAA             Context        I                   45.0000             5    0.00
exampleBAM.bam             45  AAA             Context        D                   45.0000             5    0.00
exampleBAM.bam             45  CAA             Context        I                   45.0000             5    0.00
exampleBAM.bam             45  CAA             Context        D                   45.0000             5    0.00
exampleBAM.bam             45  GAA             Context        I                   45.0000             2    0.00
exampleBAM.bam             45  GAA             Context        D                   45.0000             2    0.00
exampleBAM.bam             45  TAA             Context        I                   45.0000             6    0.00
```

GATK Base Quality Score Recalibration

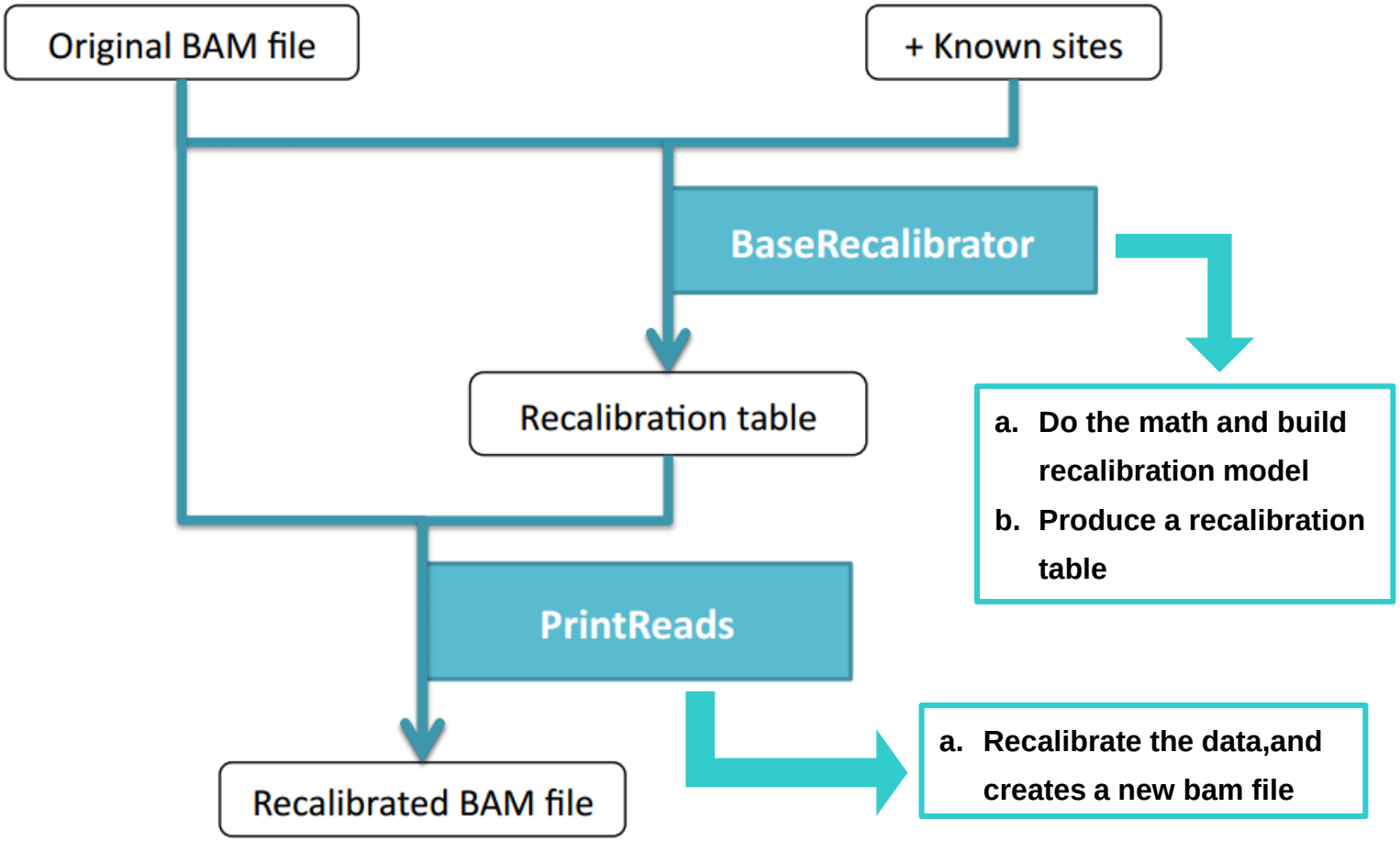
https://www.broadinstitute.org/gatk/

Base Recalibration steps/tools

a. Model the error modes and recalibrate qualities (BaseRecalibrator)

b. Write the recalibrated data to file (PrintReads)

Workflow:



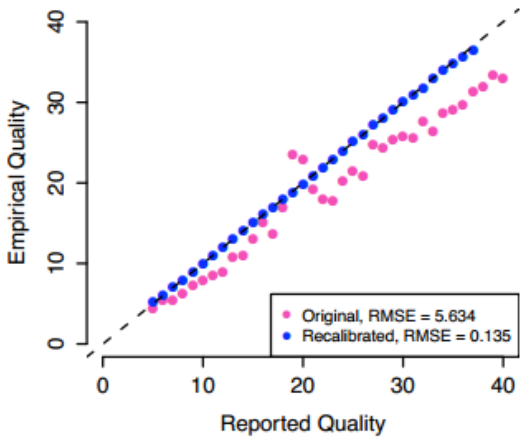
GATK Base Quality Score Recalibration

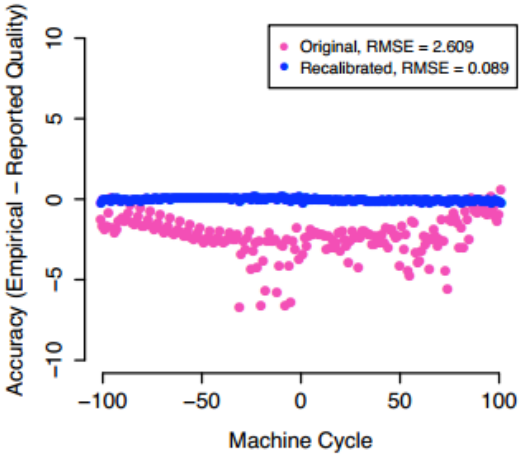
https://www.broadinstitute.org/gatk/

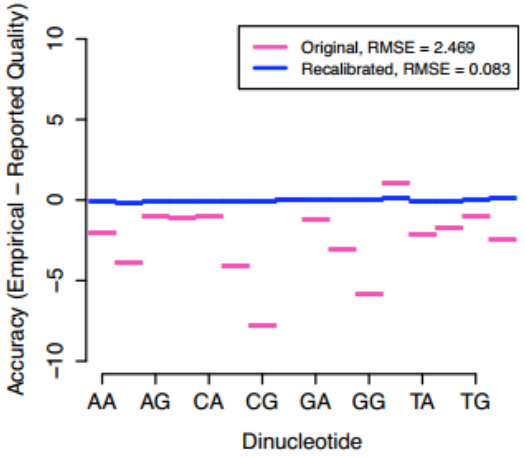
Evaluation:

a. Post-recalibration quality scores should fit the empirically-derived quality scores very well;

b. No obvious systematic biases should remain;







Notes: http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr

New with the release of the full version of **GATK 2.0** is the ability to recalibrate not only the well-known base quality scores but also base insertion and base deletion quality scores. These are per-base quantities which **estimate the probability that the next base in the read was mis-incorporated or mis-deleted** (due to slippage, for example). We've found that these new quality scores are very valuable in indel calling algorithms. In particular these new probabilities fit very naturally as the gap penalties in an HMM-based indel calling algorithms. We suspect there are many other fantastic uses for these data.

Remark:

a. What BQSR trying to do is to find the error from the sequencing machine;

b. Provide the tool sites that are known to be polymorphic, so that it is more likely to get an accurate measure of the error from the machine;

c. The critical determinant of the quality of the recalibration is the number of observed bases and mismatches in each bin. The system will not work well on a small number of aligned reads. It usually expects well **in excess of 100M bases** from a next-generation DNA sequencer **per read group**. 1B bases yields significantly better results. Keep in mind these are very general numbers to give you an idea of the range, rather than absolute thresholds. **These numbers refer to the amount of useable data, which typically would not include duplicates**.

GATK Base Quality Score Recalibration

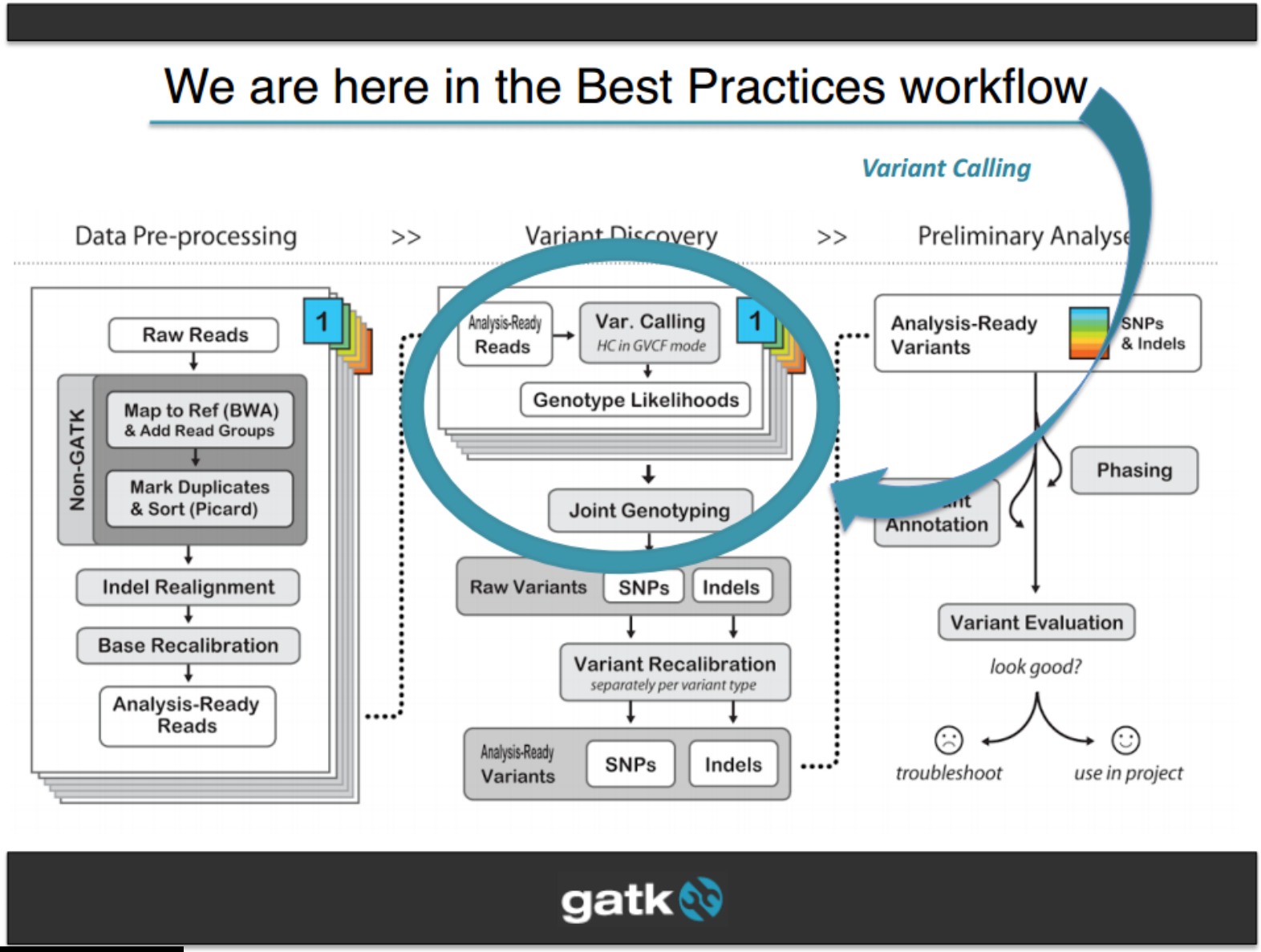
https://www.broadinstitute.org/gatk/

Command lines:

```
 java -Xmx4g -jar GenomeAnalysisTK.jar  -T BaseRecalibrator  -I my_reads.bam  \
-R resources/Homo_sapiens_assembly18.fasta  \
-knownSites bundle/hg18/dbsnp_132.hg18.vcf  \
-knownSites another/optional/setOfSitesToMask.vcf  \
-o recal_data.table

java  -Xmx2g  -jar  GenomeAnalysisTK.jar  -T  PrintReads  -R ref.fasta  \
-I realigned.bam  -BQSR recal_data.table  -o recal.bam
```

GATK Variants calling

https://www.broadinstitute.org/gatk/



Emma_JiangChongyi

GATK Variants calling

https://www.broadinstitute.org/gatk/

Likelihood

How probable is the evidence given that our hypothesis is true?

Prior

How probable was our hypothesis *before* observing the evidence?

$$P(H \mid e) = \frac{P(e \mid H)\, P(H)}{P(e)}$$

Posterior

How probable is our hypothesis given the observed evidence? (Not directly computable)

Marginal

How probable is the new evidence under all possible hypotheses?

$P(e) = \sum P(e \mid H_i)\, P(H_i)$

Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\ \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\} \Pr\{D|G_i\}}, \quad \text{[Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \quad \text{where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$\Pr\{D|H\}$ is the haploid likelihood function

Emma_JiangChongyi

Variant callers in GATK

- UnifiedGenotyper

  Call SNPs and indels separately by considering each variant locus independently

  - Accepts any ploidy
  - Pooled calling

- HaplotypeCaller

  Call SNPs, indels, and some SVs simultaneously by doing local re-assembly and considering haplotypes

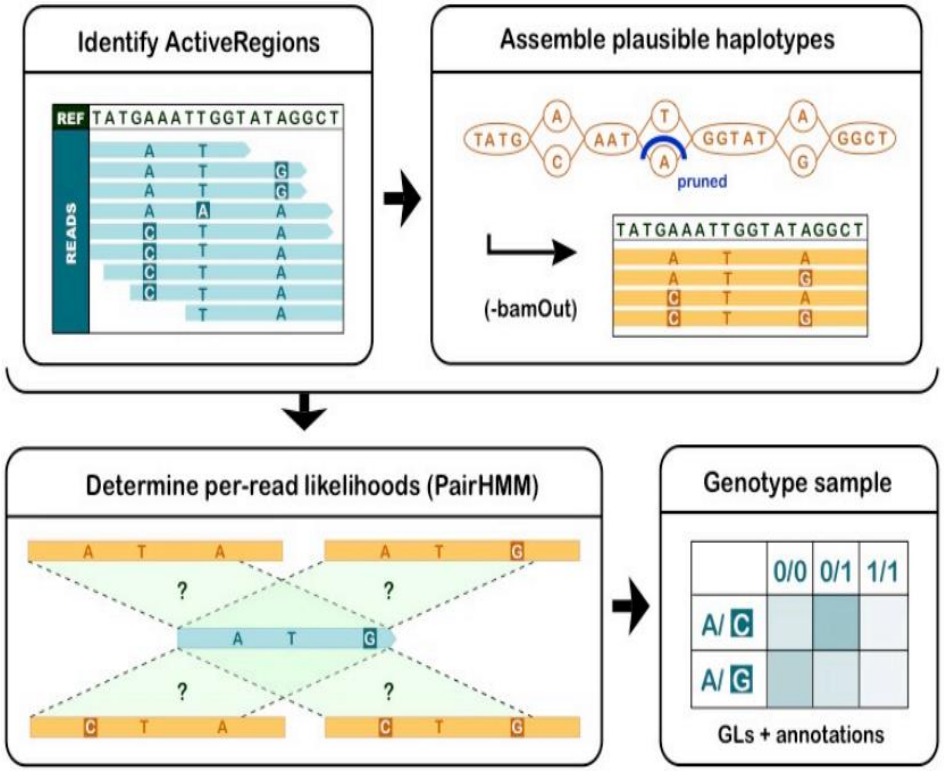
  - More accurate (esp. indels)
  - Reference confidence model
  - Replaces UG

GATK Variants calling

https://www.broadinstitute.org/gatk/guide/article?id=4148

Method Overview: Call SNPs, indels, and some SVs simultaneously by doing local re-assembly and considering haplotypes.

a. Determine if a region has **potential variation**;
b. Make **deBruijn assembly graph** of the region;
c. Paths in the graph = **potential haplotypes** to evaluate;
d. Calculate **data likelihoods** given the haplotypes (PairHMM);
e. Identify variants on most likely haplotypes;
f. Compute **allele frequency distribution** to determine most likely allele count, and emit a variant call if appropriate;
g. If emitting a variant, **assign genotype** to each sample;

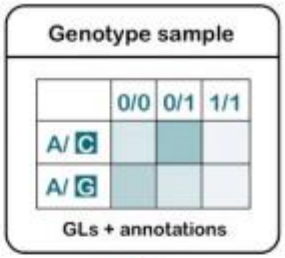


GATK Variants calling

https://www.broadinstitute.org/gatk/

Method Overview:

Summed up formally:

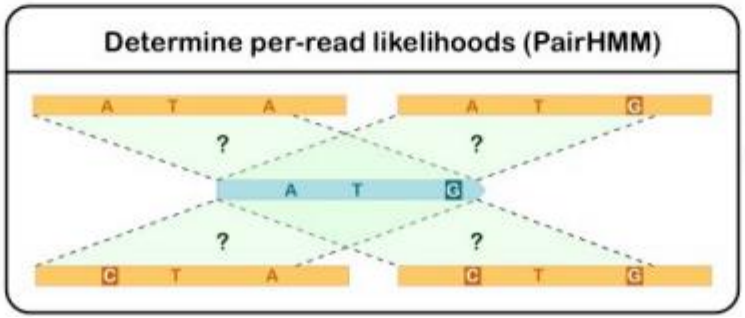


Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\ \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \quad \text{[Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \quad \text{where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$$\Pr\{D|H\} \quad \text{is} \quad \text{the haploid likelihood function}$$



GATK Variants calling

https://www.broadinstitute.org/gatk/

In the process of manual review, we found local assembly with fermi is frequently more effective than the INDEL callers, which may be because of the independence of the reference sequence, the requirement of long-range consistency and the more powerful topology-based error cleaning (Zerbino and Birney, 2008). Some difficult errors such as Figure 4 are trivial to resolve with local assembly.

14) Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics, 30: 2843-51.

GATK Variants calling

https://www.broadinstitute.org/gatk/

Command lines:

```
java  -jar GenomeAnalysisTK.jar  -T HaplotypeCaller  -R ref.fasta
-I sample1.bam [-I sample2.bam ...]  \
[--dbsnp dbSNP.vcf]  [-stand_call_conf 30]  [-stand_emit_conf 10]  [-L targets.interval_list]  \
-o output.raw.snps.indels.vcf
```

-stand_call_conf (The minimum phred-scaled confidence threshold at which variants should be called)

The minimum phred-scaled Qscore threshold to separate high confidence from low confidence calls. Only genotypes with confidence >= this threshold are emitted as called sites. A reasonable threshold is 30 for high-pass calling (this is the default).

-stand_emit_conf

The minimum phred-scaled confidence threshold at which variants should be emitted (and filtered with LowQual if less than the calling threshold).
This argument allows you to emit low quality calls as filtered records.

call是指是否对该位点进行calling，emit是指calling输出结果的表示方法；

习题

1.请问什么是低质量序列？
2.请问PCR重复是什么？它是怎么产生的？为什么要去除PCR重复？
3.请简述Fastq、BAM和VCF的格式。
4.请问碱基的Phred quality是什么？有哪些质量体系？
5.请问insert length是指什么？Bwa mem和SOAPaligner的insert length分别是怎么设置的？
6.请问BAM文件中@RG信息有什么用途？如何为BAM文件添加@RG信息？
7.请问什么是interleaved file？
8.请问soft clipping和hard clipping分别是指什么？
9. 请问报错信息如下时，应该怎么处理: "java.lang.OutOfMemoryError: Java heap space"?
10.请简述pipeup格式。
11.请问为什么处理bam文件的下游工具一般都要求先对bam文件进行sort和index？
12.请问如何将一个sample的chr1的alignment信息提取出来？
13.请简述VCF和BCF文件格式。
14.在BWA比对得到的BAM文件中，覆盖某个位点的reads在该位置的碱基都是G，但GATK的HC认为该位点有变异（例如G->C）。请问这是为什么？

Thanks