Genome Variation Detection

Jia Lei

2018.9.30

Contents

1. Tips

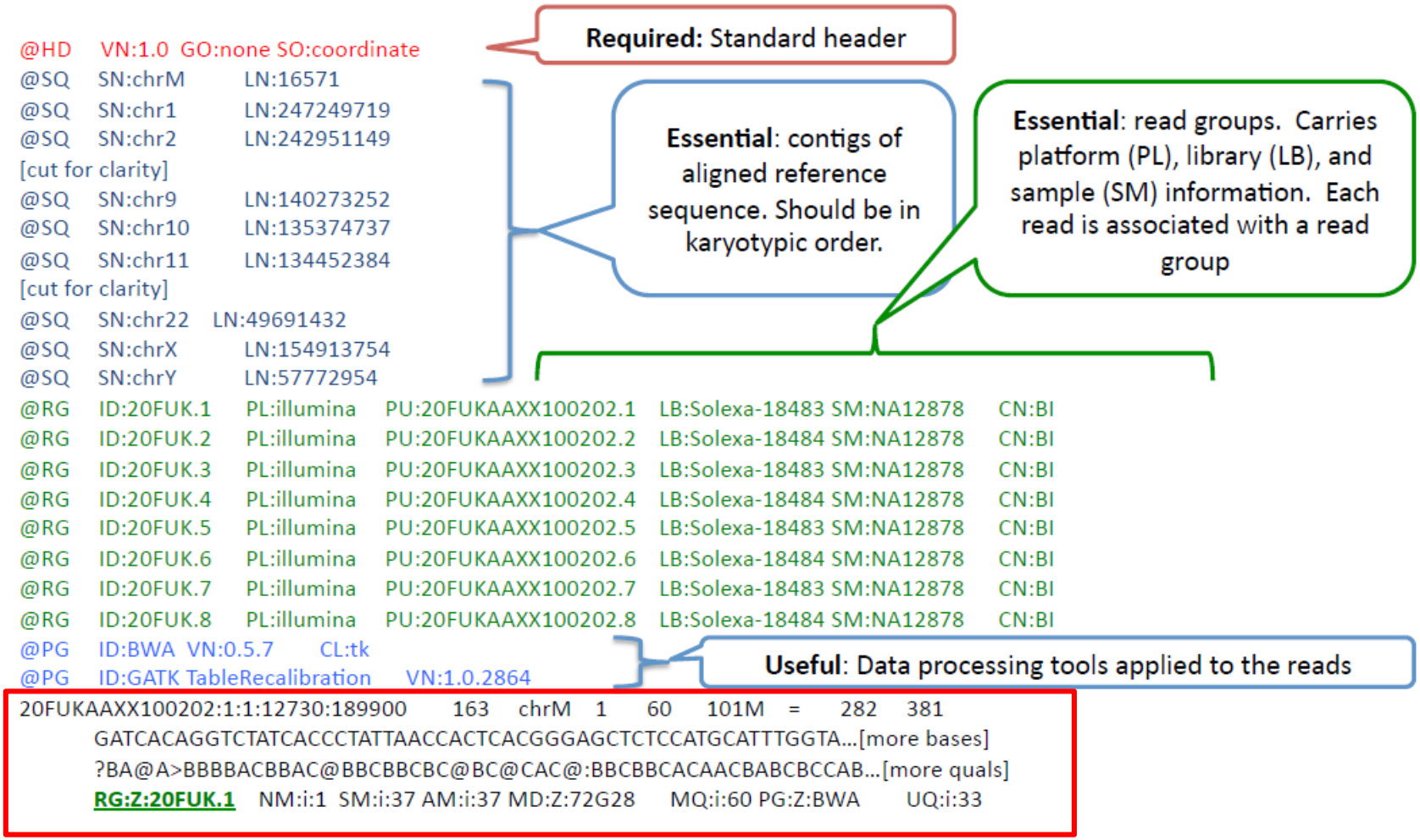
FASTQ: Raw unaligned reads

- Simple extension from traditional FASTA format.
- Each block has 4 elements (in 4 lines):
  - Sequence Name (read name, group, etc.)
  - Sequence
  - + (optional: Sequence name again)
  - Associated quality score.
- Example record:

```
@A00283:44:H5WCJDSXX:4:1101:5032:1031 1:N:0:ATCCTG
CNGAATTAACGGTCTAGCGATGGCTTCAGCTCCACTCCCATAGGCAGCACAACTGGGGTATAAGCCAAACGTCTTC
+
F#FFFFFF:FFFFFF:FFFFFFFFFFFFFFFFF:FFFFFFFFFF,FFFFFFFFFFFFFFF:FFFFFFFFFI
@A00283:44:H5WCJDSXX:4:1101:6551:1031 1:N:0:ATCCTG
ANCATGGCCCTCCCTGCAATGCGTTGAGCTCATCATCGGTGCCTCCCCATTTCTACCCGAACCACCACTGCTCACA
+
F#F,FF,FFFFFFFFFFFFFFFFF,FFFFFFFF,FFFF:F,FFFFFFFFFF,F,F:FFF:FFFFFFFF,:FI
```

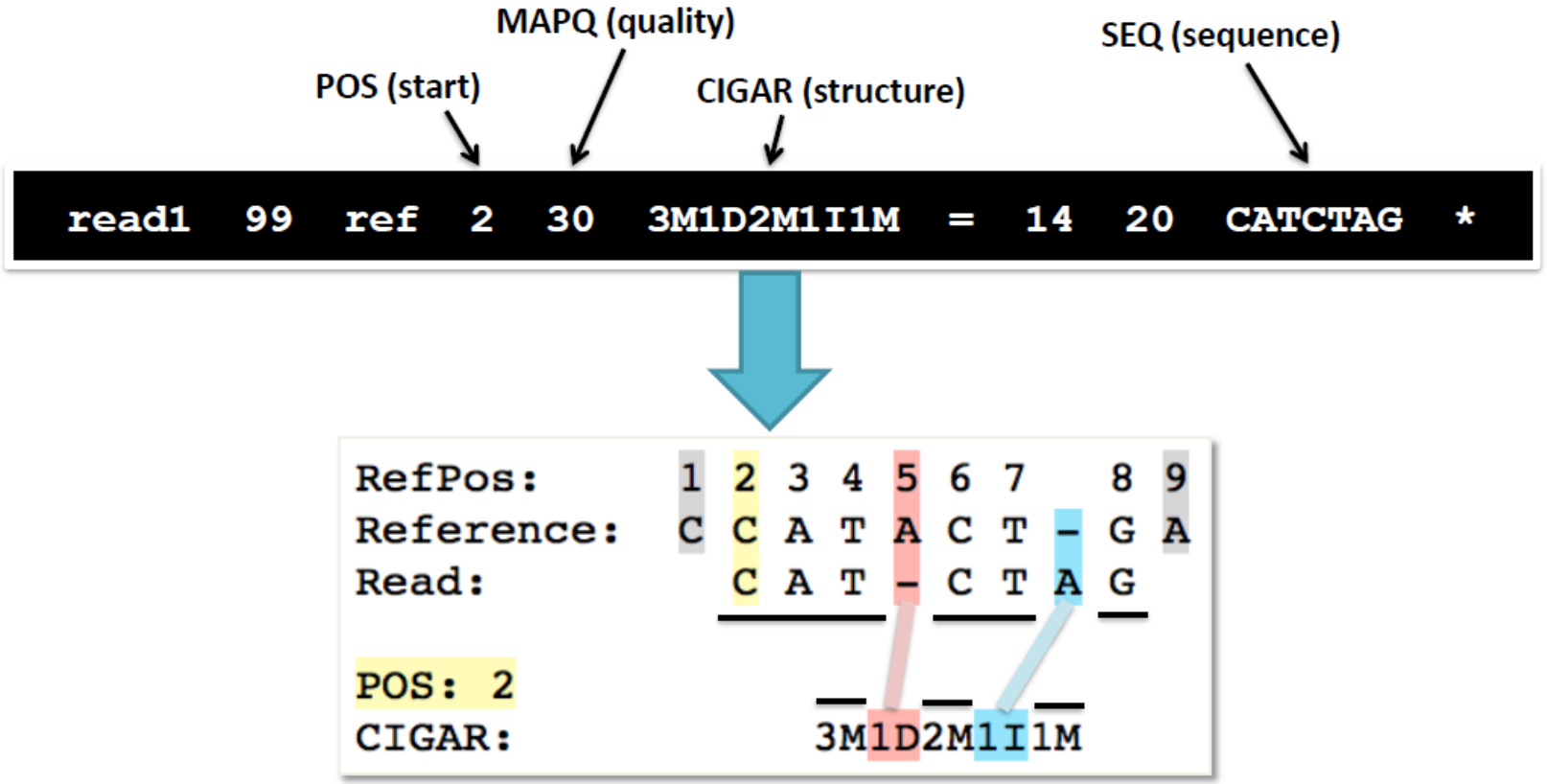
Official specification in http://maq.sourceforge.net/fastq.shtml

BAM headers: an essential part of a BAM file



Official specification in http://samtools.sourceforge.net/SAM1.pdf

Anatomy of a SAM alignment



See also:

- SAM format spec: http://samtools.github.io/hts-specs/SAMv1.pdf
- Explain SAM flags: http://broadinstitute.github.io/picard/explain-flags.html

VCF Files store variant information

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO
    FORMAT      NA00001        NA00002        NA00003
```

Header

```
20     14370   rs6054257 G       A       29    PASS    DP=14;AF=0.5;DB
    GT:GQ:DP 0|0:48:1 1|0:48:8 1/1:43:5
20     1110696 rs6040355 A       G,T     67    PASS    DP=10;AF=0.333,0.667;DB
    GT:GQ:DP 1|2:21:6 2|1:2:0   2/2:35:4
20     1230237 .         T       .       47    PASS    DP=13
    GT:GQ:DP 0|0:54:7 0|0:48:4 0/0:61:2
20     1234567 microsat1 GTCT    G,GTACT 50    PASS    DP=9
    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Variant records

Official specification in
www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41

https://blog.csdn.net/wangyiqi806643897/article/details/25423711

Bowtie2 & samtools syntax

Index the reference genome

```
bowtie2-build part_rice.fa part_rice
```

Mapping reads to reference genome

```
bowtie2 -x part_rice -1 20000_R1.fq -2 20000_R2.fq -S test.sam
```

samtools

```
samtools --help
```

http://cncbi.github.io/Bowtie2-Manual-CN/
http://www.jianshu.com/p/15f3499a6469

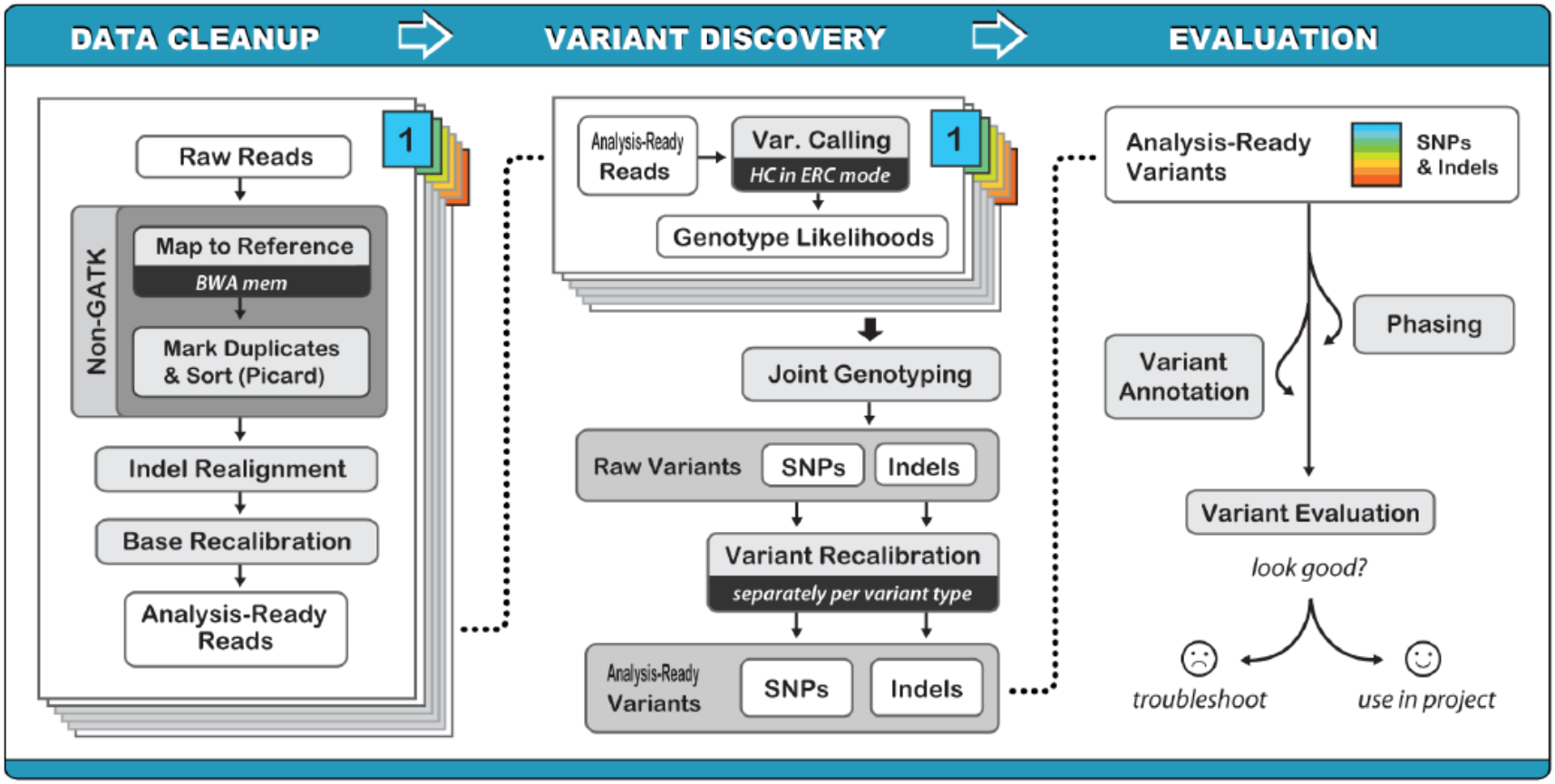
GATK command syntax

```
java –jar GenomeAnalysisTK.jar –T ToolName \
        -R reference.fasta \
        -I inputBAM.bam \
        -V inputVCF.vcf \
        -o outputs.someformat \
        -L 20:1000000-2000000
```

- Java-based command line tool (see running requirements in FAQs)
- Consult online Documentation for details about each tool!
  - Argument names and default values can change
  - Exact arguments depend on the given tool

2. SNP calling

Best Practices for Variant Discovery in DNAseq



Map to Reference

Index the reference genome

```
bowtie2-build part_rice.fa part_rice
samtools faidx part_rice.fa
samtools dict part_rice.fa -o part_rice.dict
```

Mapping reads to reference genome

```
bowtie2 -p 5 -x part_rice -1 20000_R1.fq -2 20000_R2.fq --rg-id test
--rg "PL:ILLUMINA" --rg "SM:test" -S test.sam
```

http://cncbi.github.io/Bowtie2-Manual-CN/

Data Processing

Sam to bam

```
samtools view -bS test.sam > test.bam
samtools view -bS test.sam -o test.bam
```

Sort bam file

```
samtools sort -O BAM -o test.sorted.bam test.bam
```

Index bam file

```
samtools index test.sorted.bam
```

manual-cn
http://www.chenlianfu.com/?p=1399

Remove PCR duplicates

Remove duplicates

```
samtools rmdup -sS test.sorted.bam test.rmdup.bam
```

Index bam file

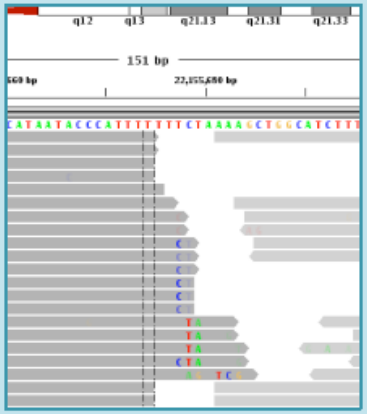
```
samtools index test.rmdup.bam
```

manual-cn
http://www.chenlianfu.com/?p=1399

Indel Realignment steps/tools

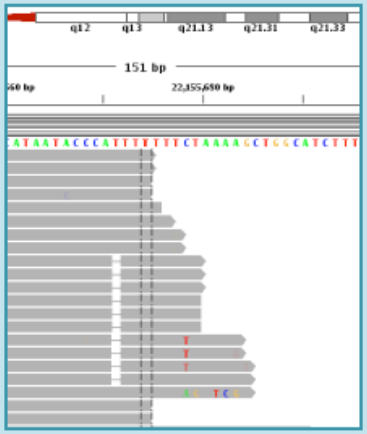
- Identify what regions need to be realigned

→ **RealignerTargetCreator**



- Perform the actual realignment

→ **IndelRealigner**



Why realign around indels?

- InDels in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches
- These artifactual mismatches can harm base quality recalibration and variant detection (unless a sophisticated caller like the Haplotype Caller is used)

☑ **Realignment around indels helps improve the accuracy of several of the downstream processing steps.**

RealignerTargetCreator

- Pre-processing step to find intervals that may need realignment

```
java -Djava.io.tmpdir=./tmp \
       -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
        -R part_rice.fa \
        -I test.rmdup.bam \
        -o test.realn.intervals
```

- Input BAM file not necessary if processing only at known indels
- Using a list of known indels will both speed up processing and improve accuracy, but is not required

IndelRealigner

- Attempts realignment at RealignerTargetCreator target intervals

```
java -Djava.io.tmpdir=./tmp \
     -jar GenomeAnalysisTK.jar -T IndelRealigner \
     -R part_rice.fa \
     -I test.rmdup.bam \
     -targetIntervals test.realn.intervals \
     -o part_rice.realigned.bam
```

- Must use same input file(s) used in RealignerTargetCreator step
- Processing options
  - Only at known indels: much faster, accurate for ~90-95% of indels
  - At indels seen in the original BAM alignments: the recommended mode
  - Using full Smith-Waterman realignment: most accurate, but heavy computational cost and not really necessary with the new techs

UnifiedGenotyper

Call SNPs and indels on a per-locus basis

```
java -Djava.io.tmpdir=./tmp  \
        -jar GenomeAnalysisTK.jar \
        -R part_rice.fa  \
        -T UnifiedGenotyper \
        -I part_rice.realigned.bam  \
        -o part_rice.raw.vcf  \
        --genotype_likelihoods_model BOTH \
        -rf BadCigar \
        -stand_call_conf 30
```

Base Recalibration steps/tools

- Model the error modes and recalibrate qualities

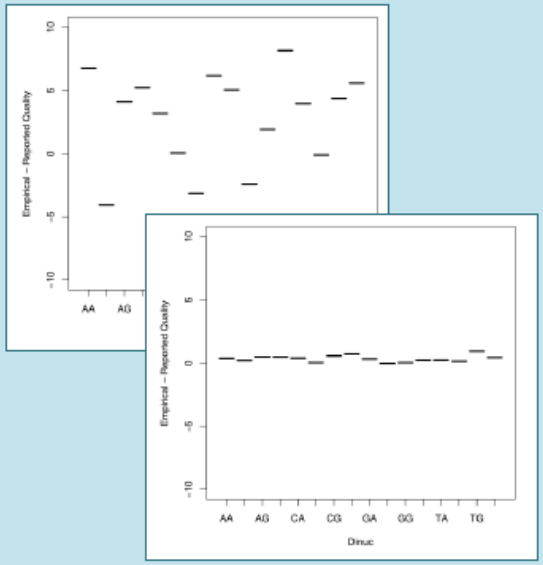
  → **BaseRecalibrator**

- Write the recalibrated data to file

  → **PrintReads**

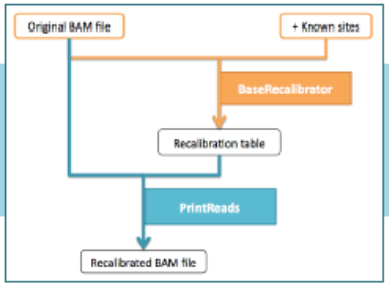
- Make before/ after plots

  → **AnalyzeCovariates**



TOOL TIPS

BaseRecalibrator

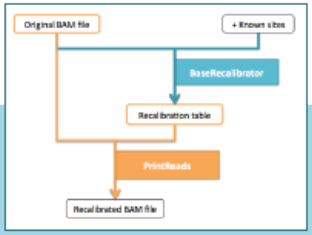


- Builds recalibration model

```
java -Djava.io.tmpdir=./tmp \
        -jar GenomeAnalysisTK.jar \
       -T BaseRecalibrator \
       -R part_rice.fa \
       -I part_rice.realigned.bam \
      -o part_rice.recal_data.grp \
      -knownSites part_rice.raw.vcf
```

TOOL TIPS

Print Reads



- General-use tool co-opted with –BQSR flag and fed a recalibration report

```
java -Djava.io.tmpdir=./tmp \
        -jar GenomeAnalysisTK.jar \
        -T PrintReads \
        -R part_rice.fa \
        -I part_rice.realigned.bam \
        -o part_rice.recal.bam \
        -BQSR part_rice.recal_data.grp
```

- Creates a new bam file using the input table generated previously which has exquisitely accurate base substitution, insertion, and deletion quality scores

- Original qualities retained with OQ tag

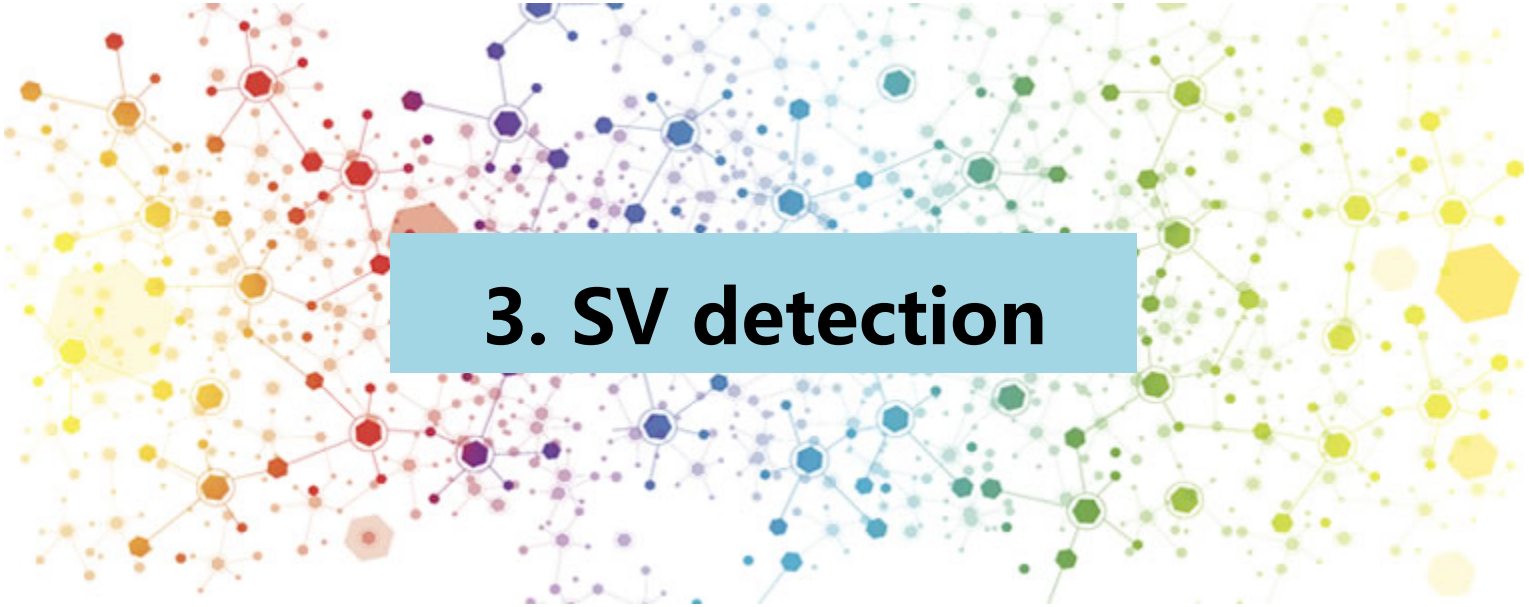
UnifiedGenotyper

Call SNPs and indels on a per-locus basis

```
java -Djava.io.tmpdir=./tmp \
        -jar GenomeAnalysisTK.jar \
        -R part_rice.fa \
        -T UnifiedGenotyper \
        -I part_rice.recal.bam \
        -o part_rice.vcf \
        -stand_call_conf 30  \
        -glm SNP \
        -allowPotentiallyMisencodedQuals
```

Further reading

http://www.broadinstitute.org/gatk/guide/best-practices

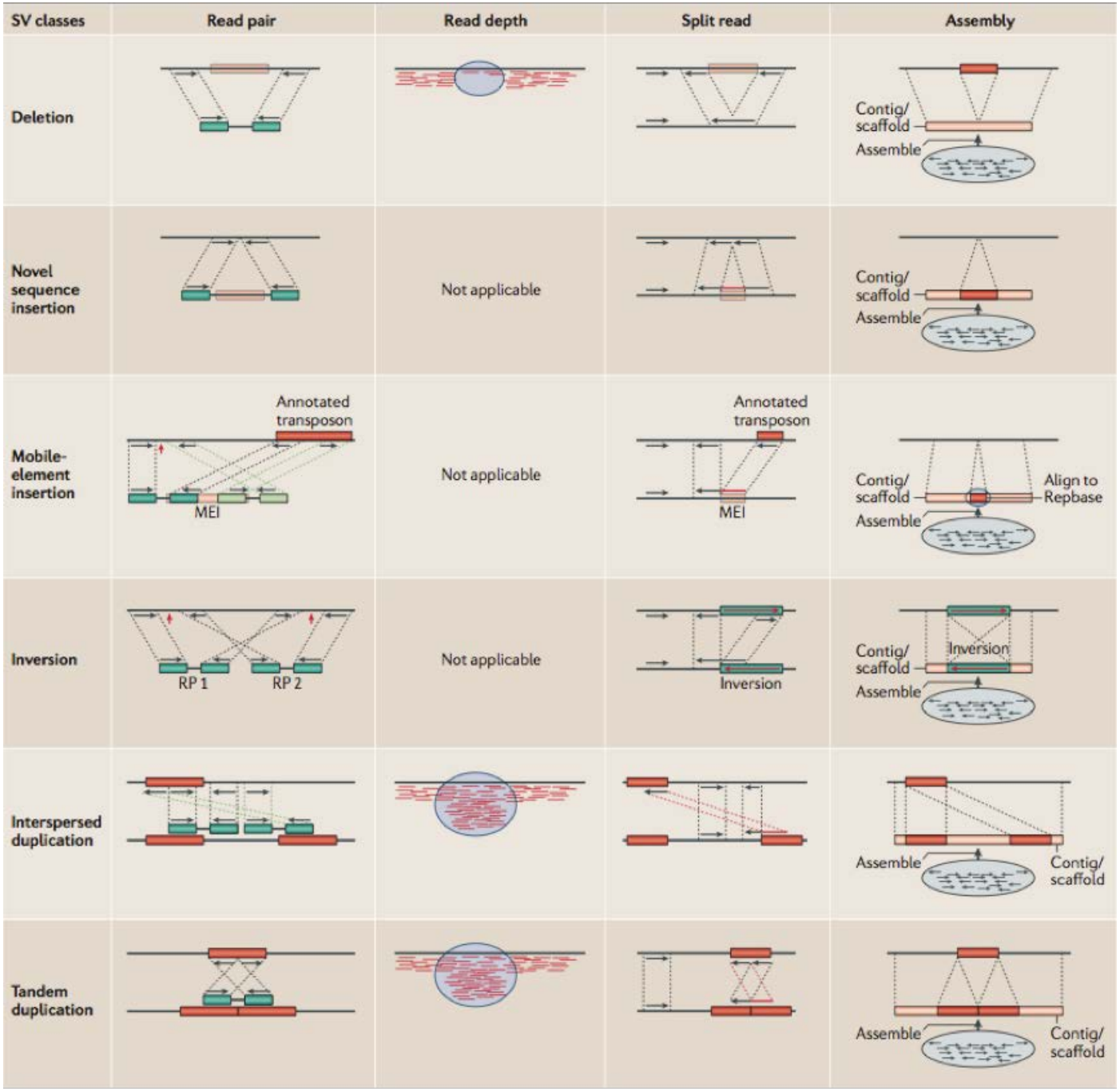


3. SV detection

SV detection strategy based on NGS data

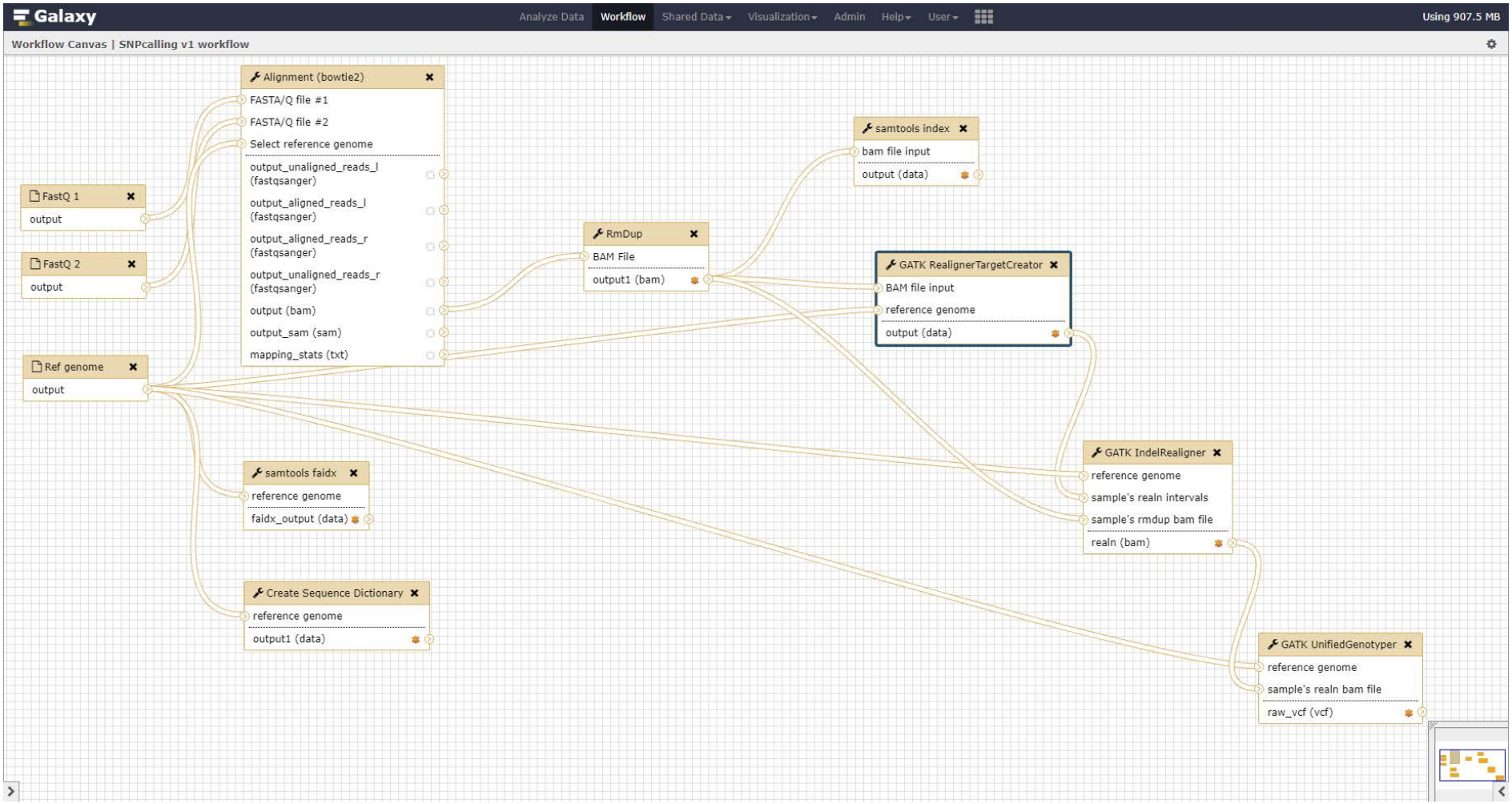
- Read Pair, RP
- Split Read, SR
- Read Depth, RD
- *de novo* Assembly, AS

https://www.jianshu.com/p/4c8e109f0e6a



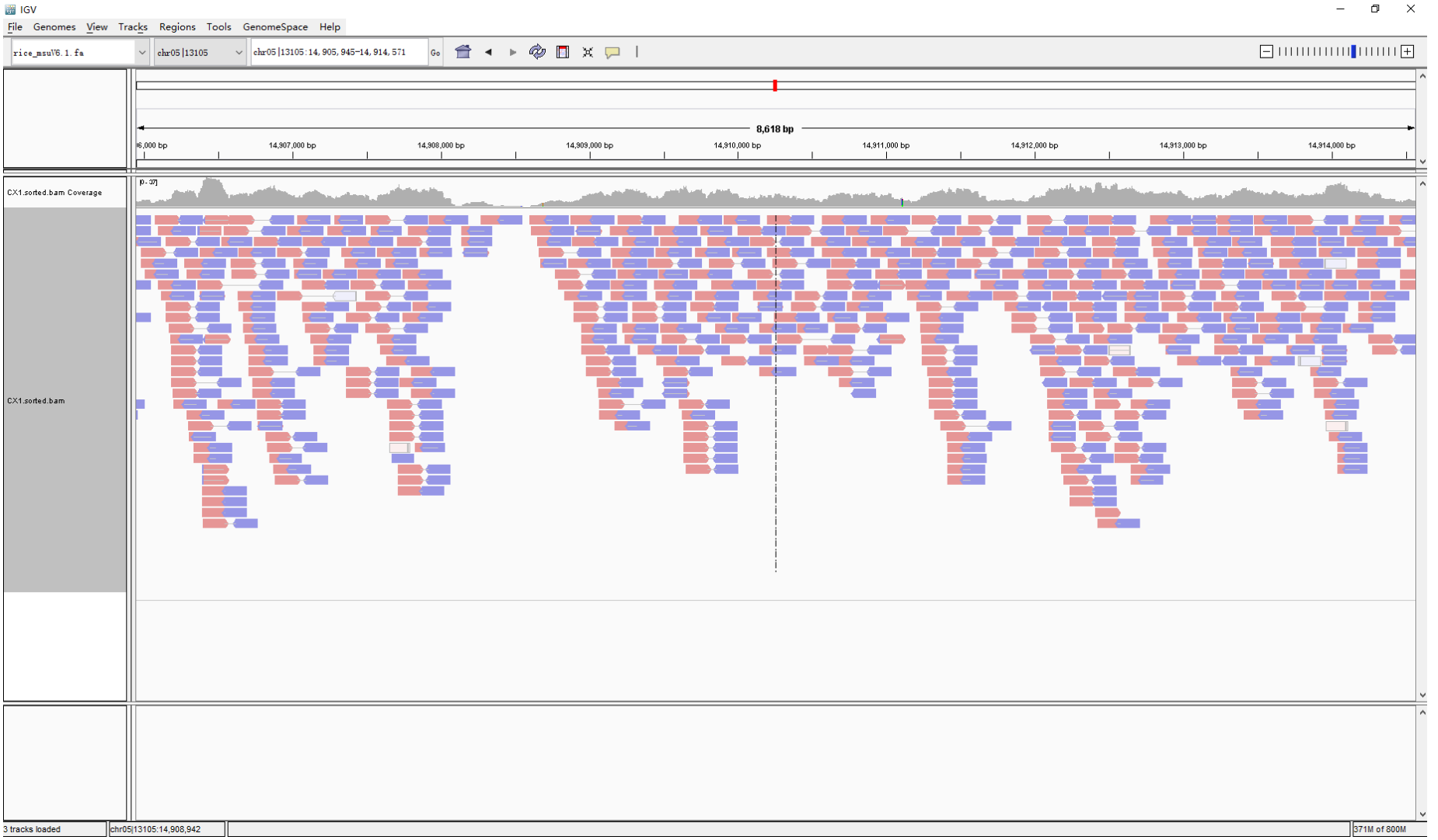
https://www.jianshu.com/p/4c8e109f0e6a


4. More



https://usegalaxy.org/

IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
rice_msu6.1.fa
chr05|13105
chr05|13105:14,905,945-14,914,571
Go
8,618 bp
14,907,000 bp
14,908,000 bp
14,909,000 bp
14,910,000 bp
14,911,000 bp
14,912,000 bp
14,913,000 bp
14,914,000 bp
CX1.sorted.bam Coverage
CX1.sorted.bam
3 tracks loaded
chr05|13105:14,908,942
371M of 800M


Thanks!