Analysing re-sequencing samples

Malin Larsson

Malin.larsson@scilifelab.se

WABI / SciLifeLab

Re-sequencing

Reference genome assembly

```
...GTGCGTAGACTGCTAGATCGAAGA...
```

Re-sequencing

IND 1
GTAGACT
AGATCGG
GCGTAGT

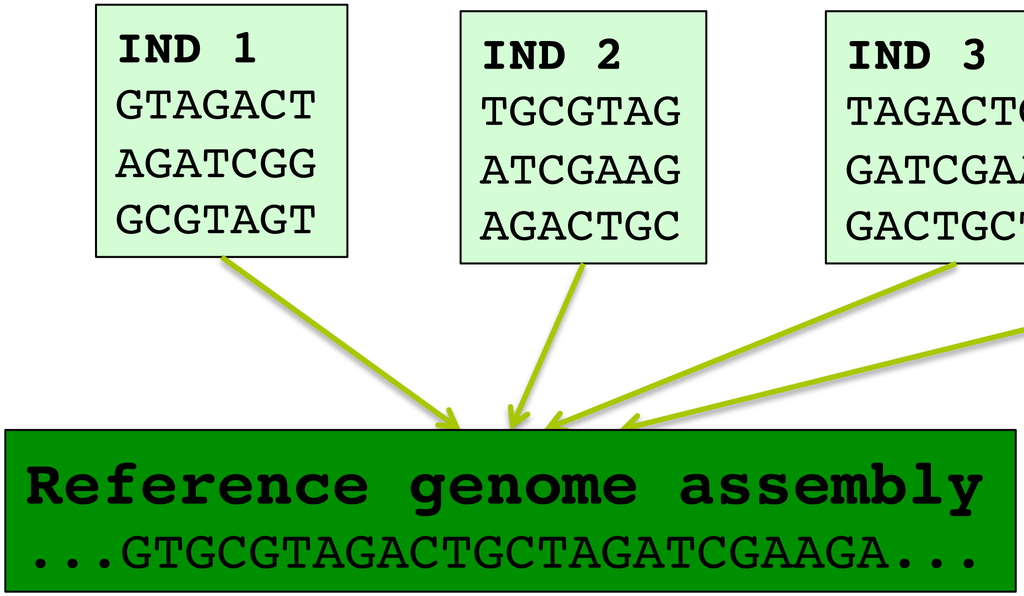
IND 2
TGCGTAG
ATCGAAG
AGACTGC

IND 3
TAGACTG
GATCGAA
GACTGCT

IND 4
AGATCGA
GTAGACT
GATCGAA

Reference genome assembly
...GTGCGTAGACTGCTAGATCGAAGA...

Re-sequencing



IND 1
GTAGACT
AGATCGG
GCGTAGT

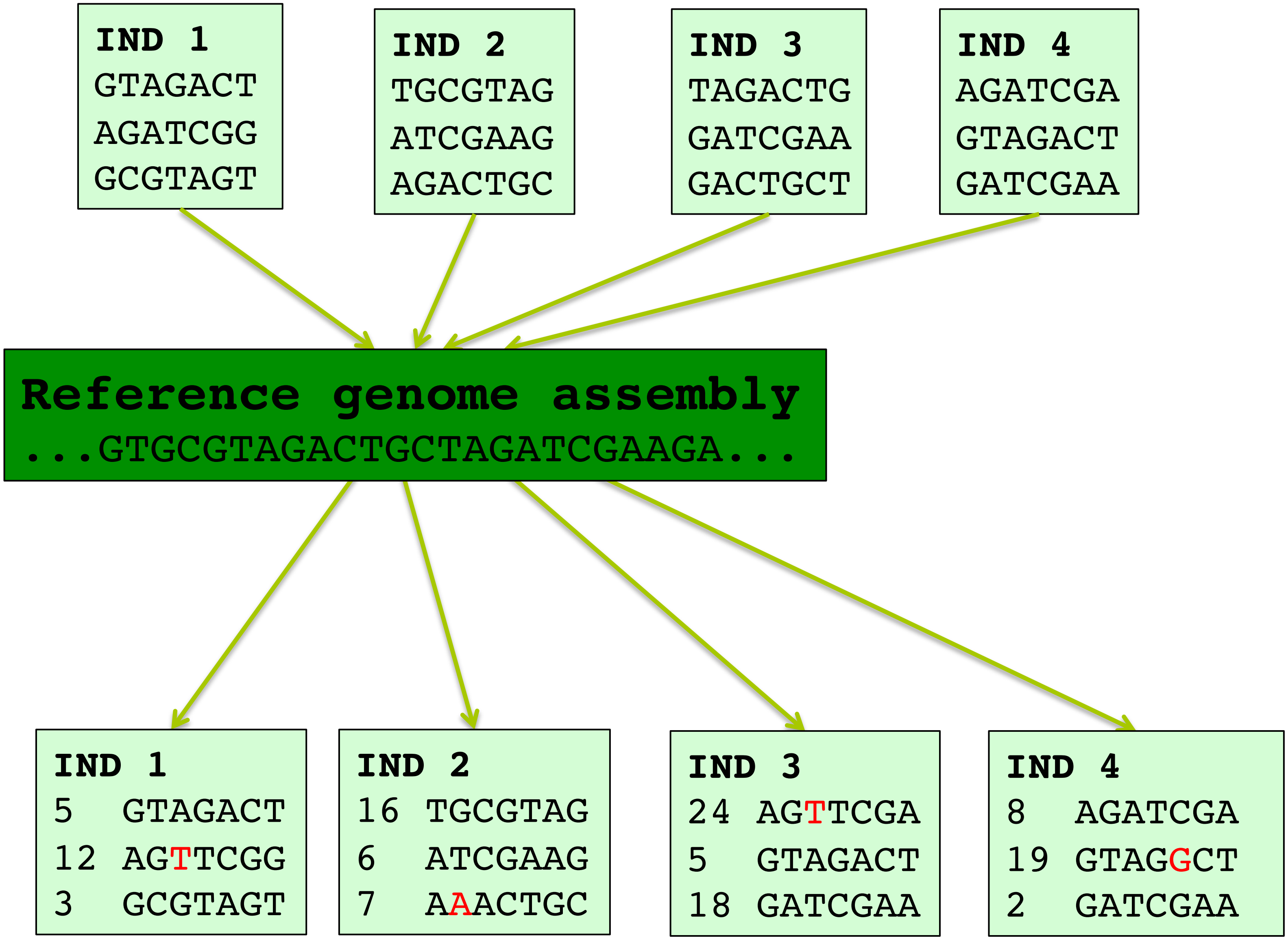
IND 2
TGCGTAG
ATCGAAG
AGACTGC

IND 3
TAGACTG
GATCGAA
GACTGCT

IND 4
AGATCGA
GTAGACT
GATCGAA

Reference genome assembly
...GTGCGTAGACTGCTAGATCGAAGA...

Re-sequencing



IND 1
GTAGACT
AGATCGG
GCGTAGT

IND 2
TGCGTAG
ATCGAAG
AGACTGC

IND 3
TAGACTG
GATCGAA
GACTGCT

IND 4
AGATCGA
GTAGACT
GATCGAA

Reference genome assembly
...GTGCGTAGACTGCTAGATCGAAGA...

IND 1
5 GTAGACT
12 AGTTCGG
3 GCGTAGT

IND 2
16 TGCGTAG
6 ATCGAAG
7 AAACTGC

IND 3
24 AGTTCGA
5 GTAGACT
18 GATCGAA

IND 4
8 AGATCGA
19 GTAGGCT
2 GATCGAA

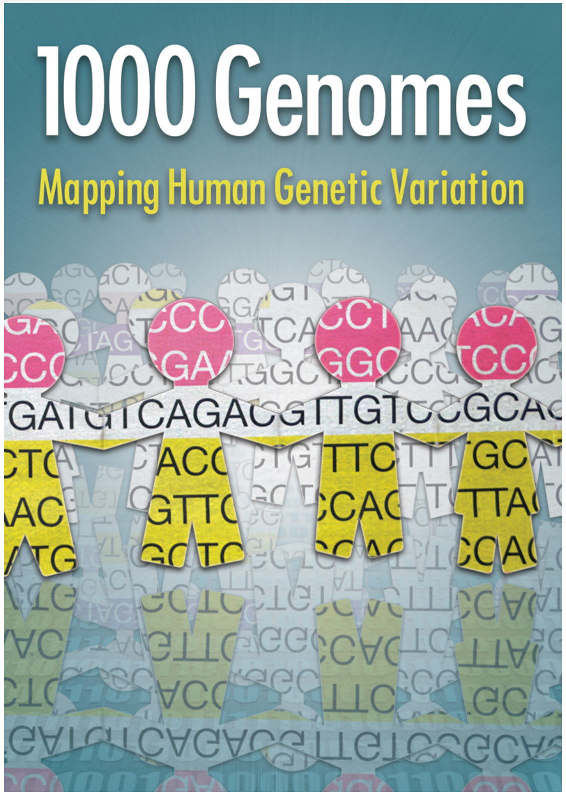
Rare variants in human

SciLifeLab

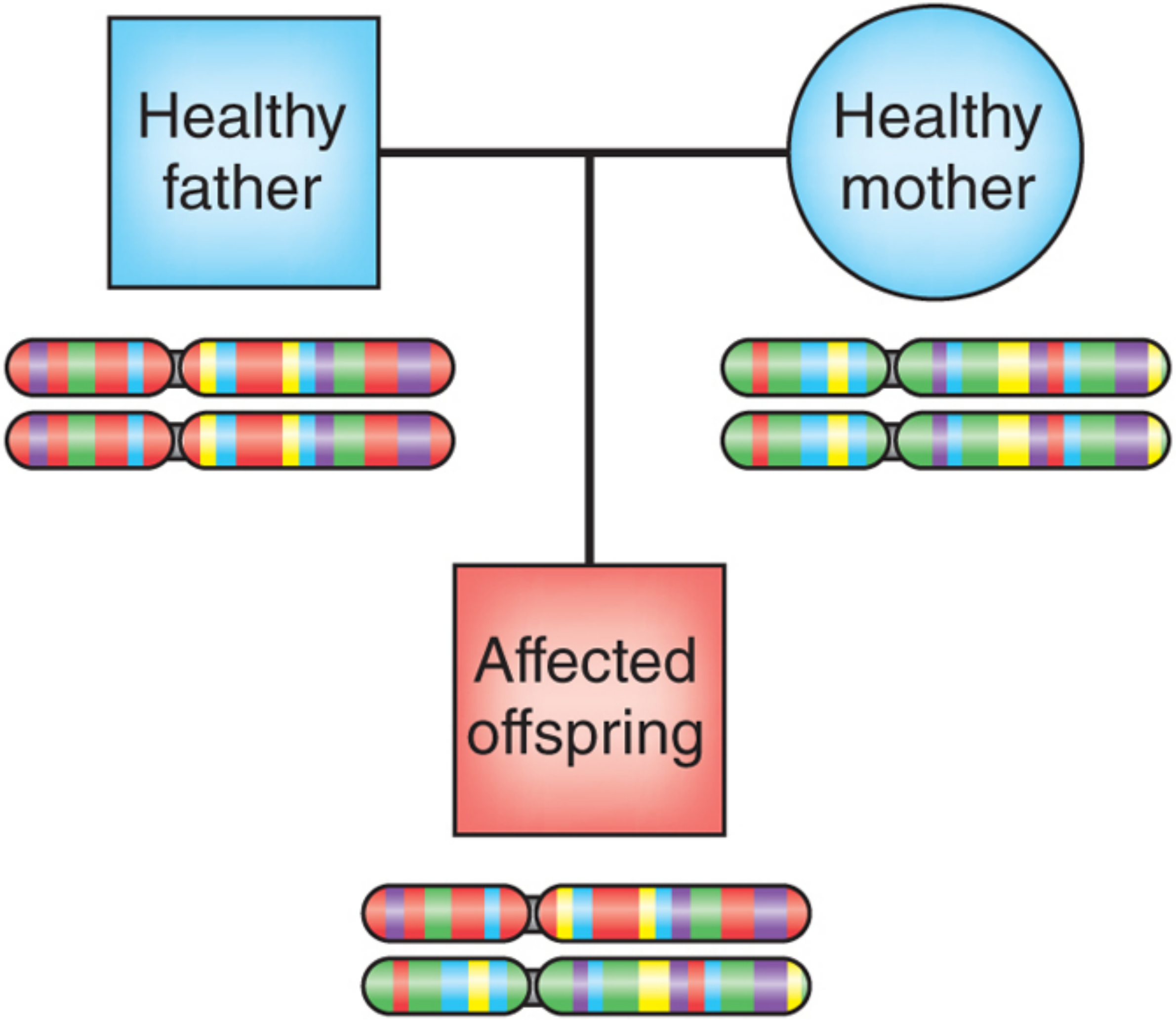
deCODE genetics

UK
10K

Rare Genetic Variants in Health and Disease

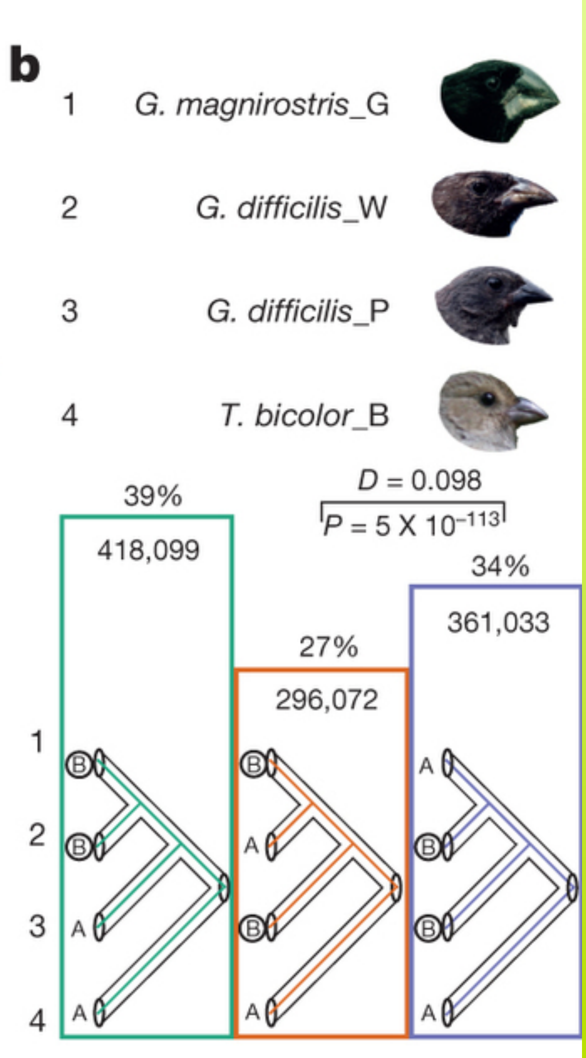


Exome sequencing in trios to detect *de novo* coding variants



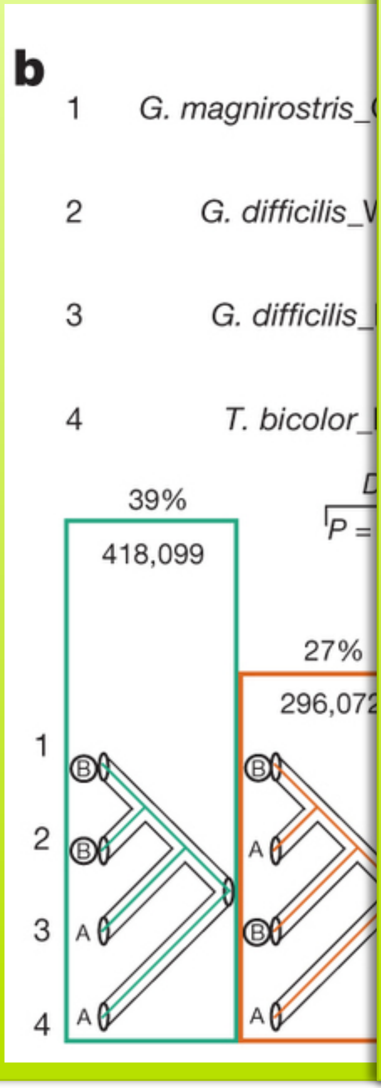
Population genetics – speciation, adaptive evolution

Darwin Finches

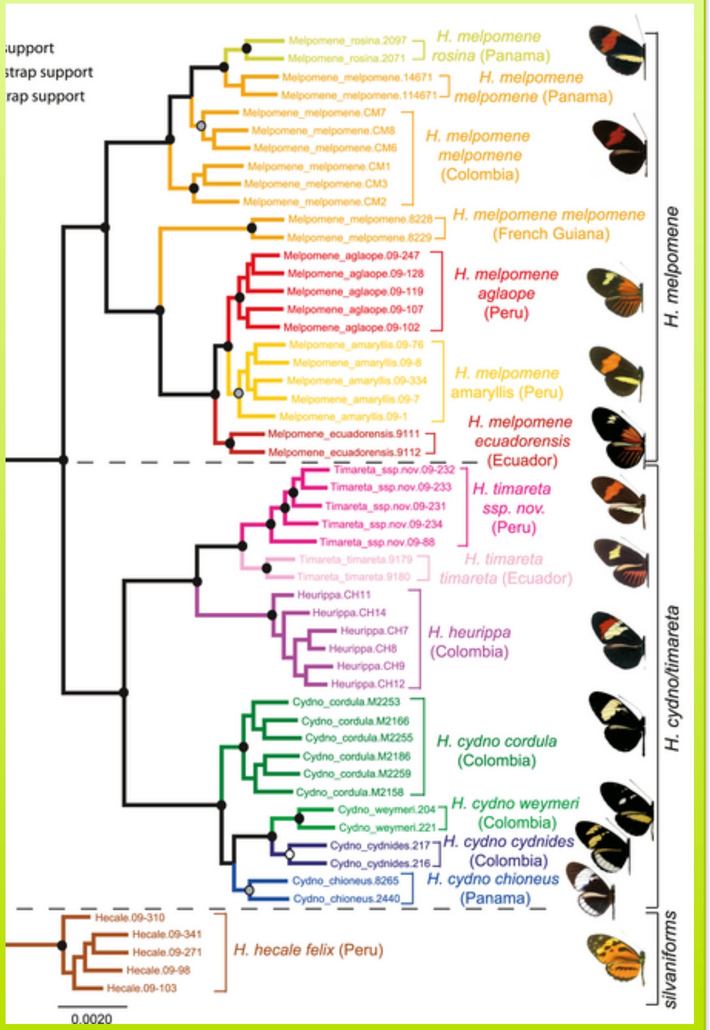


Population genetics – speciation, adaptive evolution

Darwin Finches

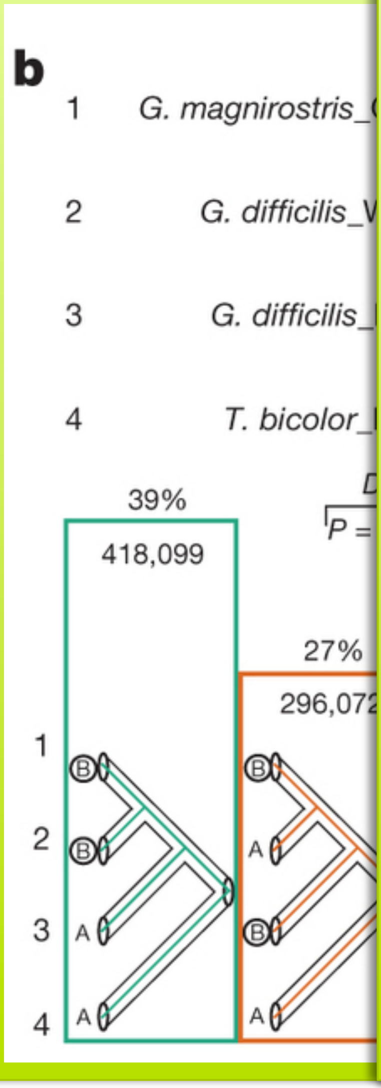


Heliconius Butterflies

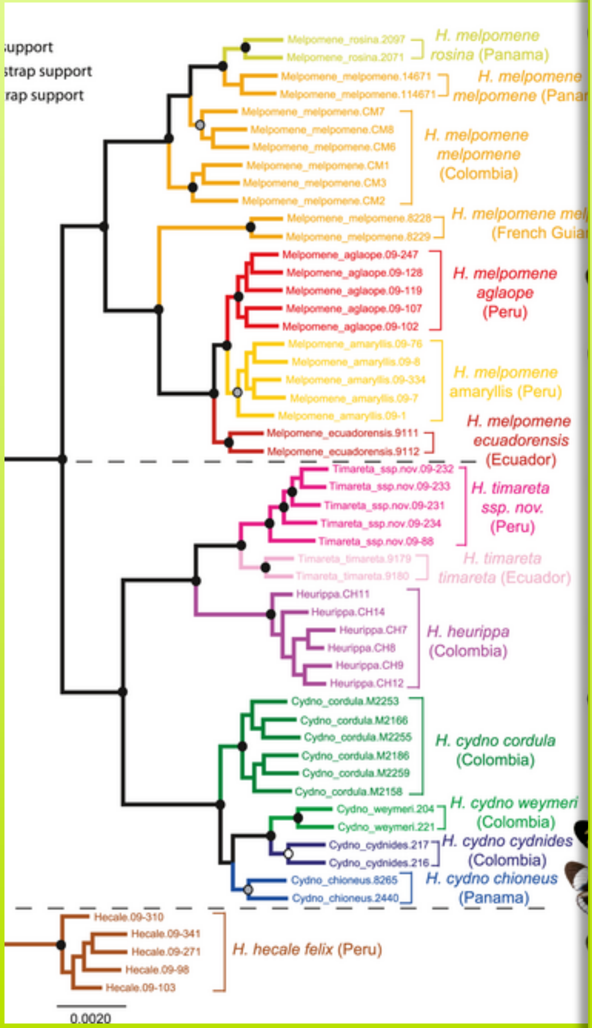


Population genetics – speciation, adaptive evolution

Darwin Finches

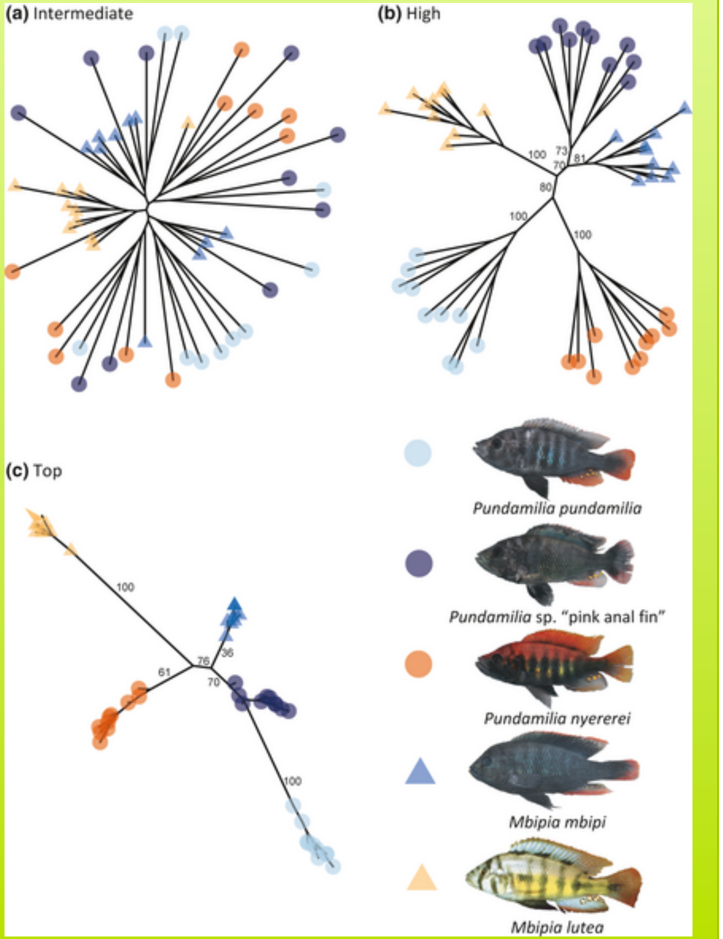


Heliconius Butterflies

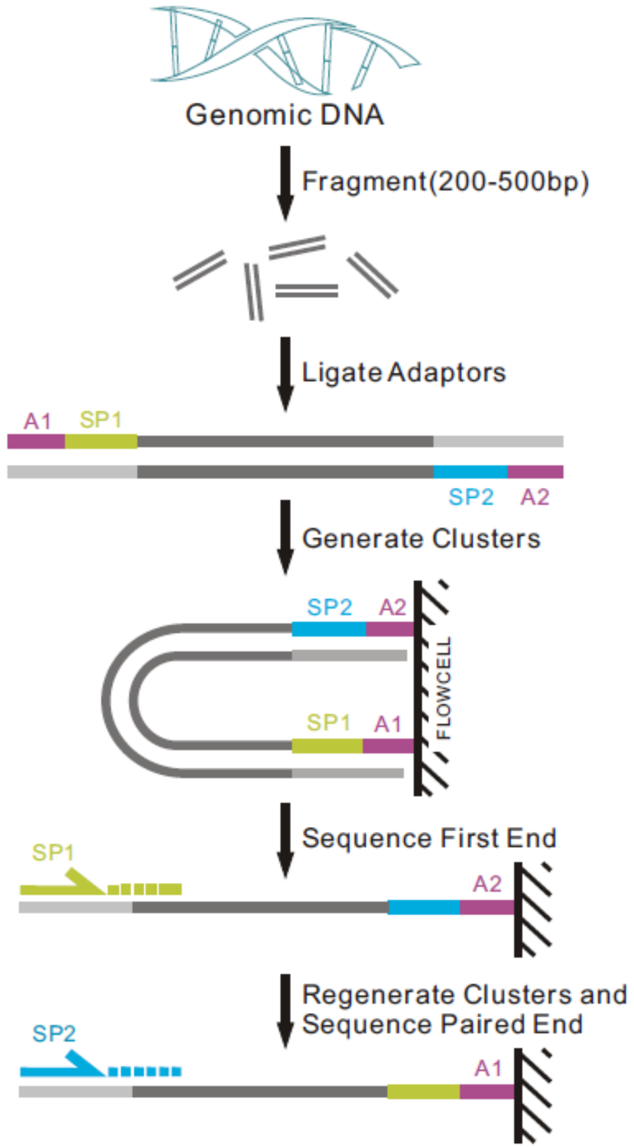


support
strap support
rap support
H. melpomene rosina (Panama)
H. melpomene melpomene (Panama)
H. melpomene melpomene (Colombia)
H. melpomene melpomene (French Guiana)
H. melpomene aglaope (Peru)
H. melpomene amaryllis (Peru)
H. melpomene ecuadorensis (Ecuador)
H. timareta ssp. nov. (Peru)
H. timareta timareta (Ecuador)
H. heurippa (Colombia)
H. cydno cordula (Colombia)
H. cydno weymeri (Colombia)
H. cydno cydnides (Colombia)
H. cydno chioneus (Panama)
H. hecale felix (Peru)
0.0020

Lake Victoria cechlid fishes



Paired end sequencing



Pair-end reads

- Two .fastq files containing the reads are created
- The order in the files are identical and naming of reads are the same with the exception of the end
- The naming of reads is changing and depends on software version

ID_R1_001.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFHHHHHGJJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 2:N:0:ATCACG
CTTCGTCCACTTTCATTATTCCTTTCATACATG
CTCTCCGGTTTAGGGTACTCTTGACCTGGCCTT
TTTTCAAGACGTCCCTGACTTGATCTTGAAACG
+
CCCFFFFFHHHHHJJJJIJJJJJJJJJJJJJJJ
JJJJJJJIJIJGIJHBGHHIIIJIJJJJJJJJI
JJJHFFFFFFDDDDDDDDDDDDDDDEDCCDDDD
```

Pair-end reads

- Two .fastq files containing the reads are created
- The order in the files are identical and naming of reads are the same with the exception of the end
- The naming of reads is changing and depends on software version

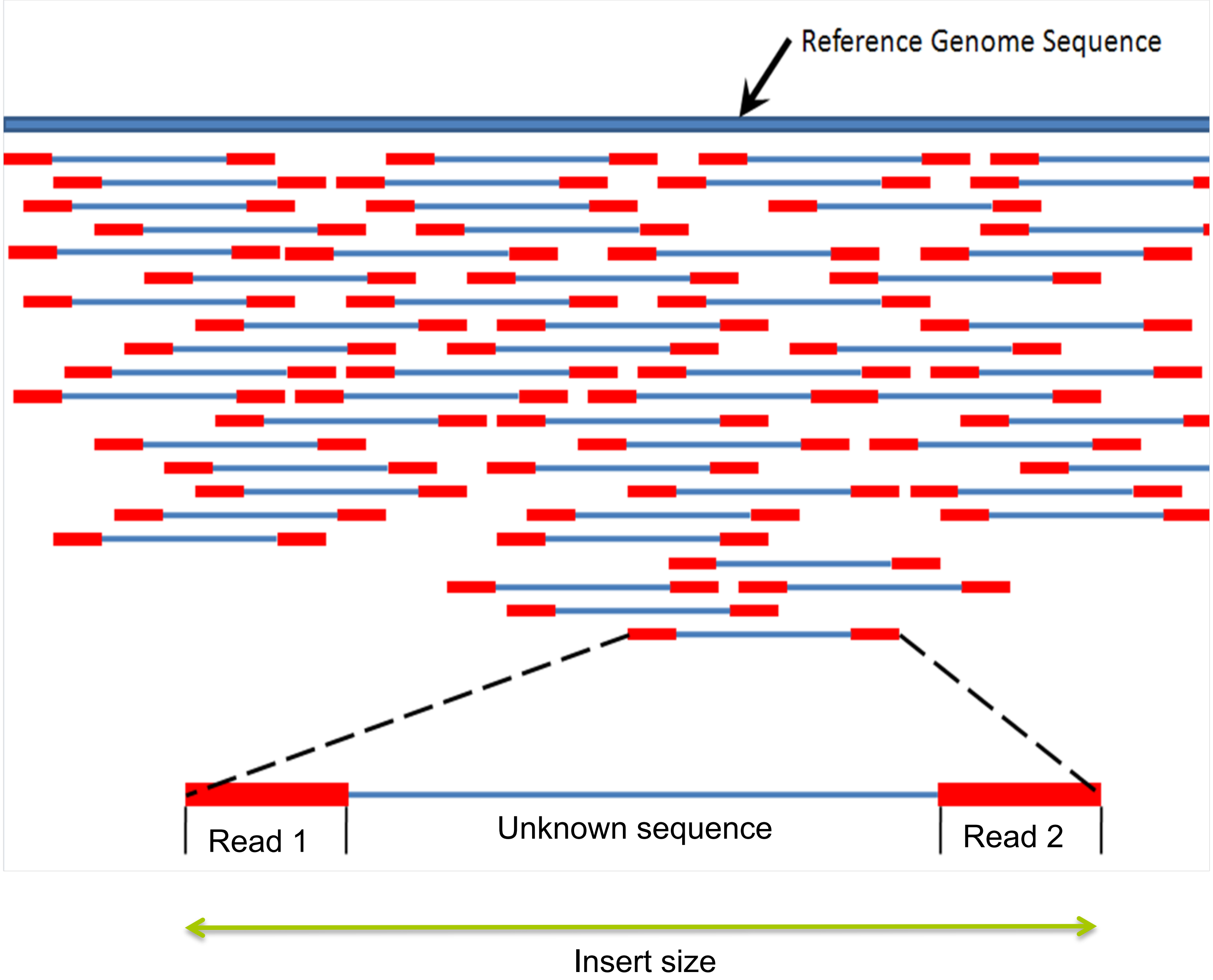
ID_R1_001.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFHHHHHGJJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 2:N:0:ATCACG
CTTCGTCCACTTTCATTATTCCTTTCATACATG
CTCTCCGGTTTAGGGTACTCTTGACCTGGCCTT
TTTTCAAGACGTCCCTGACTTGATCTTGAAACG
+
CCCFFFFFHHHHHJJJJIJJJJJJJJJJJJJJJ
JJJJJJJIJIJGIJHBGHHIIIJIJJJJJJJJI
JJJHFFFFFFDDDDDDDDDDDDDDDEDCCDDDD
```

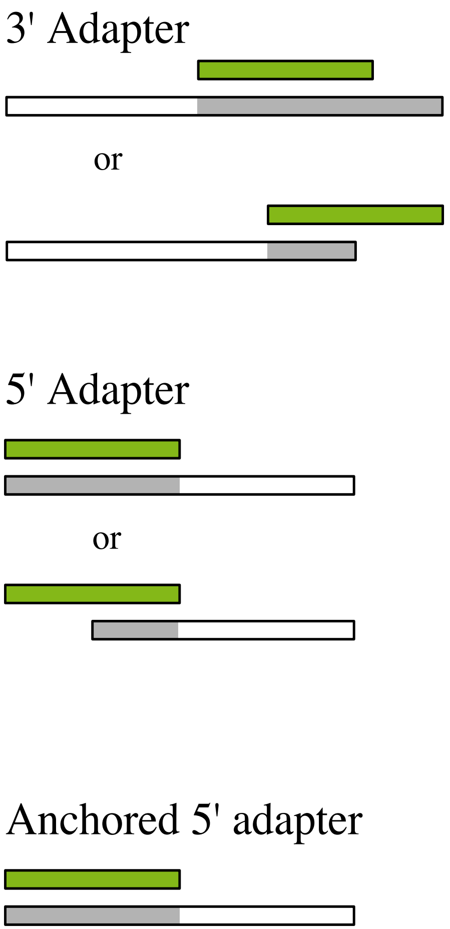
Paired end sequencing



Adapter trimming

```
Module load cutadapt
```

When the adaptor has been read in sequencing, it is present in reads and needs to be removed prior to mapping



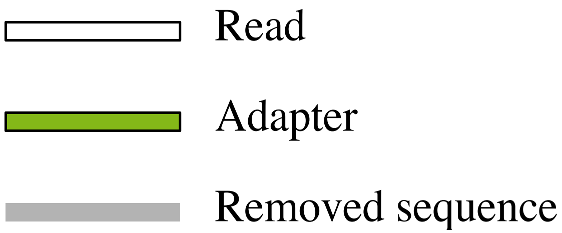
3' Adapter

or

5' Adapter

or

Anchored 5' adapter



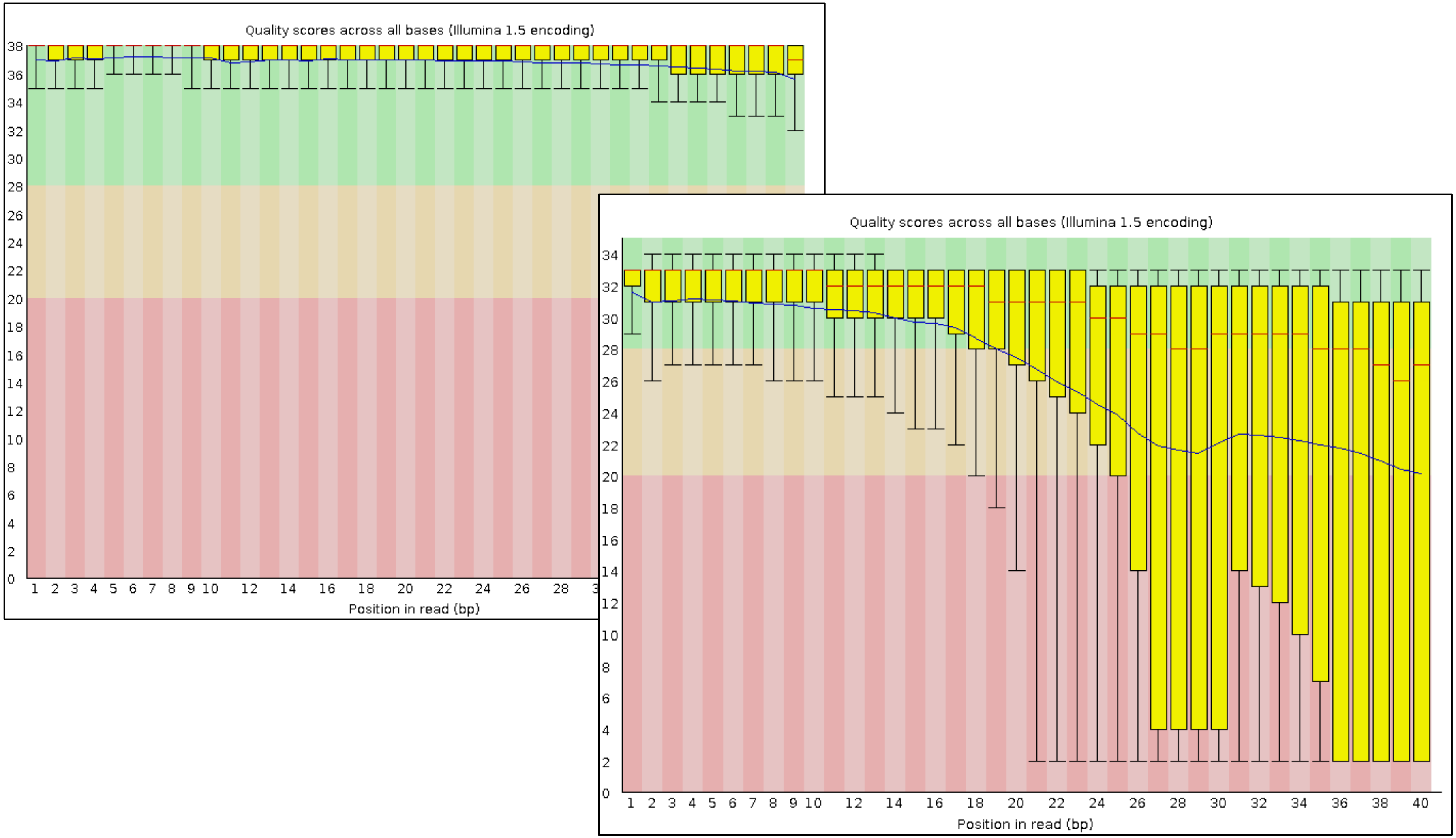
▭ Read

🟩 Adapter

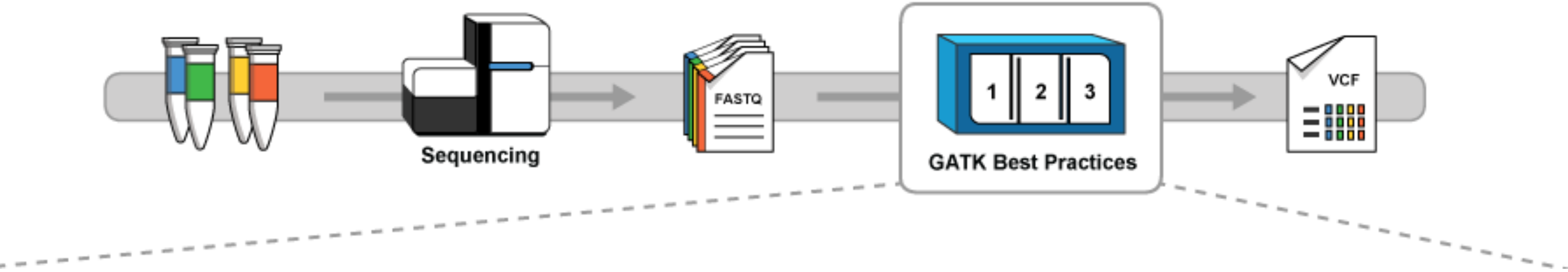
▨ Removed sequence

Basic quality control - FASTQC

```
Module load FastQC
```



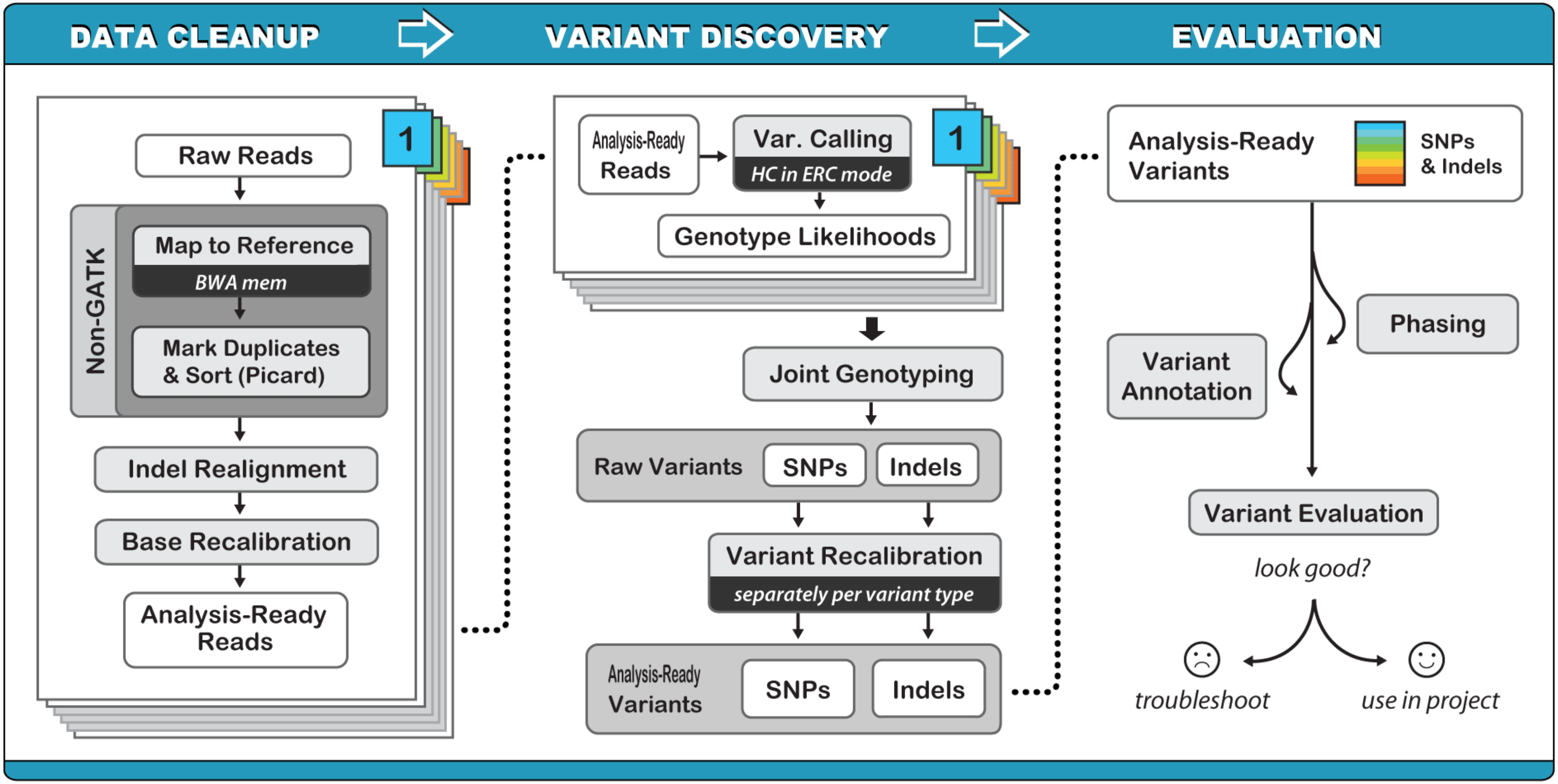
Genome Analysis Tool Kit (GATK)



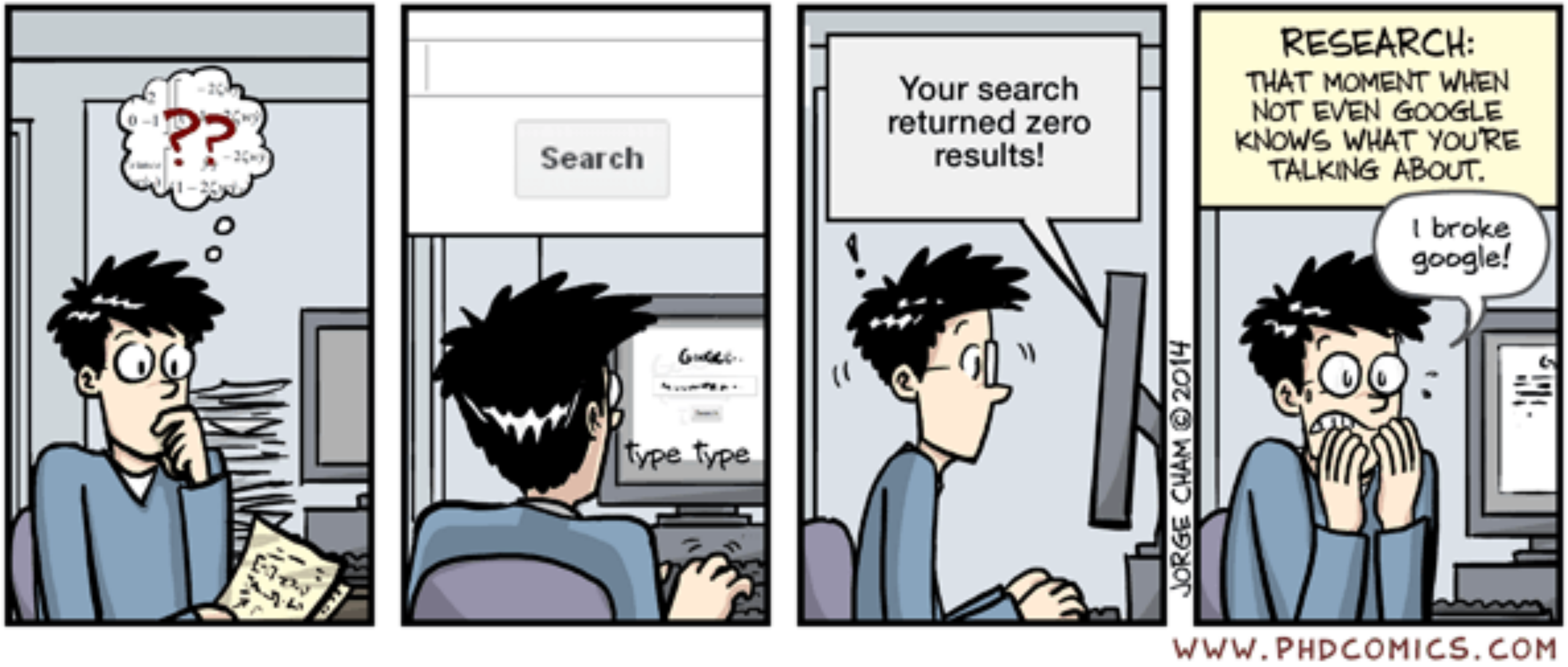
Mapping Alignment refinement Variant discovery Callset refinement

GATK

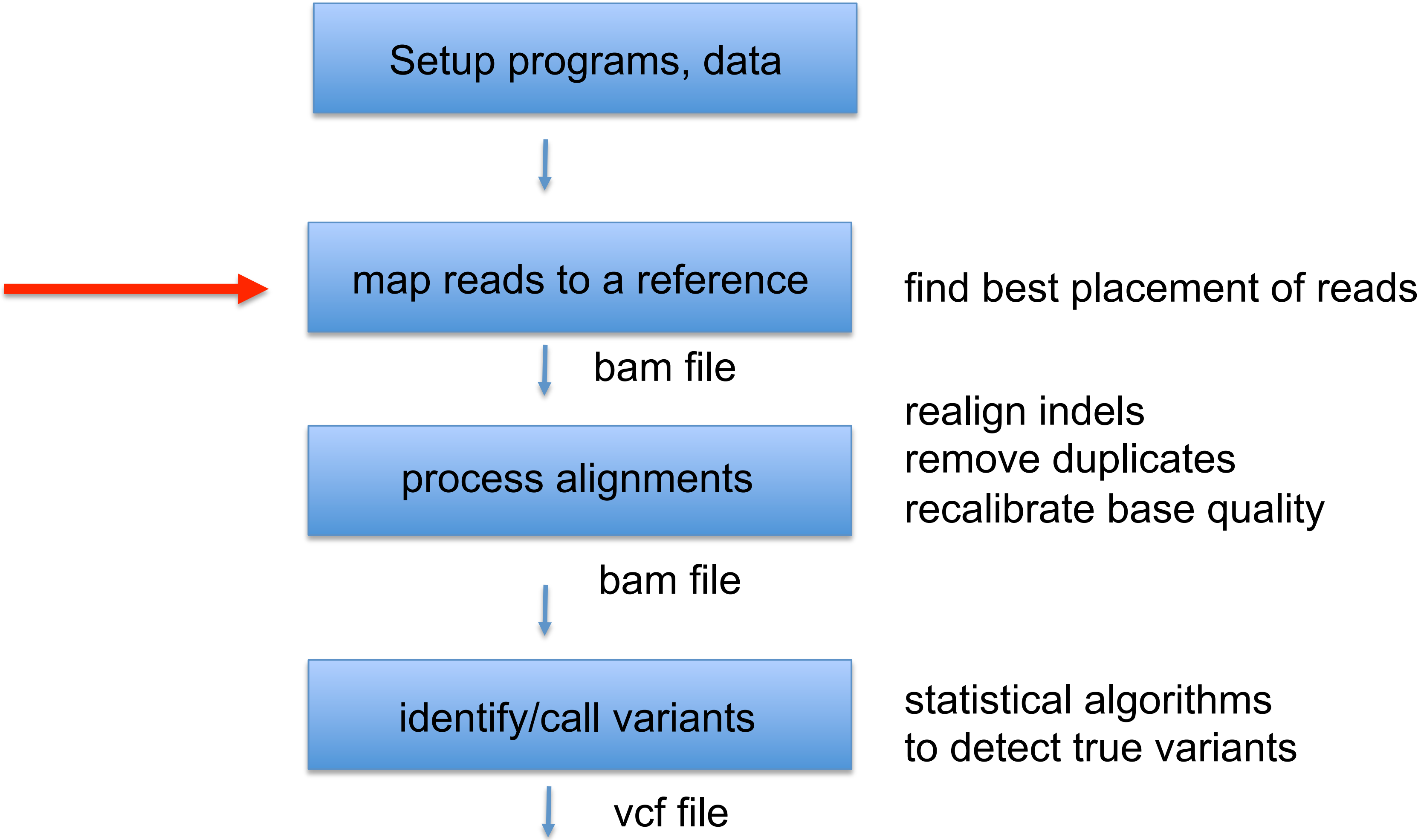
SciLifeLab



When in doubt, google it!


??
Search
type type
Your search returned zero results!
RESEARCH:
THAT MOMENT WHEN NOT EVEN GOOGLE KNOWS WHAT YOU'RE TALKING ABOUT.
I broke google!
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

Steps in resequencing analysis



Setup programs, data

map reads to a reference

find best placement of reads

bam file

process alignments

realign indels
remove duplicates
recalibrate base quality

bam file

identify/call variants

statistical algorithms
to detect true variants

vcf file

Mapping to reference genome

brute force

```
TCGATCC
x
GACCTCATCGATCCCACTG
```

brute force

```
 TCGATCC
 x
GACCTCATCGATCCCACTG
```

brute force

```
  TCGATCC
  x
GACCTCATCGATCCCACTG
```

brute force

```
   TCGATCC
   x
GACCTCATCGATCCCACTG
```

brute force

```
    TCGATCC
    ||x
GACCTCATCGATCCCACTG
```

brute force

```
     TCGATCC
     x
GACCTCATCGATCCCACTG
```

brute force

```
      TCGATCC
      x
GACCTCATCGATCCCACTG
```

brute force

```
       TCGATCC
       |||||||
GACCTCATCGATCCCACTG
```

hash tables

build an index of the reference sequence for fast access

seed length 7

```
0    5    10   15

GACCTCATCGATCCCACTG
GACCTCA             →    chromosome 1, pos 0
 ACCTCAT            →    chromosome 1, pos 1
  CCTCATC           →    chromosome 1, pos 2
   CTCATCG          →    chromosome 1, pos 3
    TCATCGA         →    chromosome 1, pos 4
     CATCGAT        →    chromosome 1, pos 5
      ATCGATC       →    chromosome 1, pos 6
       TCGATCC      →    chromosome 1, pos 7
        CGATCCC     →    chromosome 1, pos 8
         GATCCCA    →    chromosome 1, pos 9
```

hash tables

build an index of the reference sequence for fast access

TCGATCC ?

```
0    5    10   15

GACCTCATCGATCCCACTG
GACCTCA             →    chromosome 1, pos 0
 ACCTCAT            →    chromosome 1, pos 1
  CCTCATC           →    chromosome 1, pos 2
   CTCATCG          →    chromosome 1, pos 3
    TCATCGA         →    chromosome 1, pos 4
     CATCGAT        →    chromosome 1, pos 5
      ATCGATC       →    chromosome 1, pos 6
       TCGATCC      →    chromosome 1, pos 7
        CGATCCC     →    chromosome 1, pos 8
         GATCCCA    →    chromosome 1, pos 9
```

hash tables

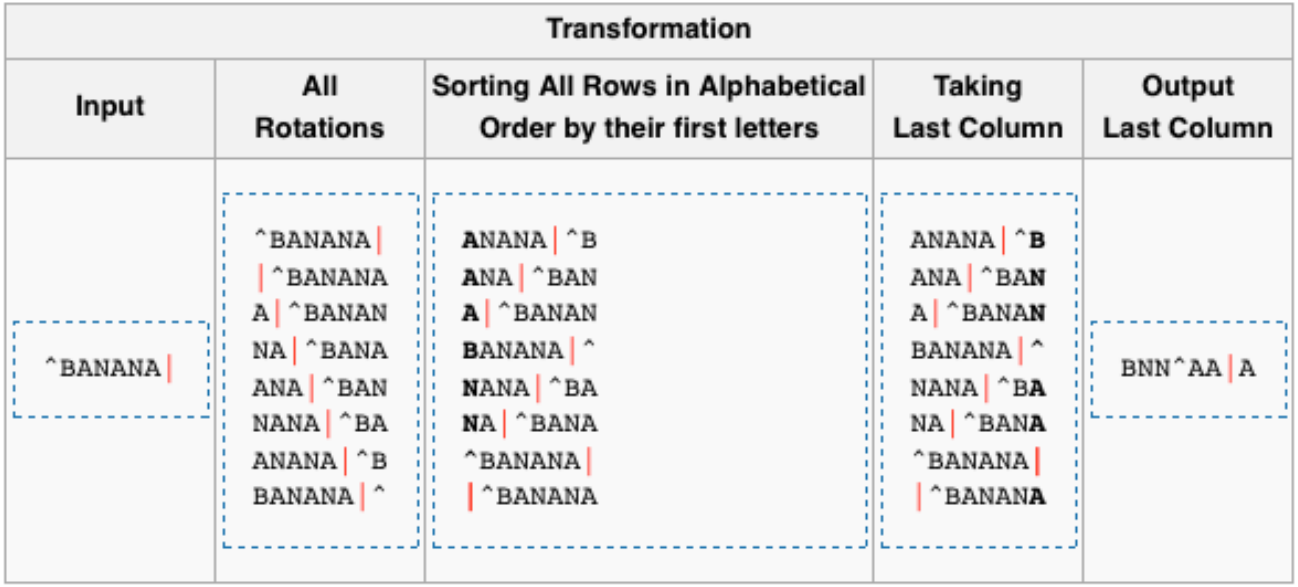
build an index of the reference sequence for fast access

```
TCGATCC = chromosome 1, pos 7

0    5    10   15

GACCTCATCGATCCCACTG
GACCTCA             →    chromosome 1, pos 0
 ACCTCAT            →    chromosome 1, pos 1
  CCTCATC           →    chromosome 1, pos 2
   CTCATCG          →    chromosome 1, pos 3
    TCATCGA         →    chromosome 1, pos 4
     CATCGAT        →    chromosome 1, pos 5
      ATCGATC       →    chromosome 1, pos 6
       TCGATCC      →    chromosome 1, pos 7
        CGATCCC     →    chromosome 1, pos 8
         GATCCCA    →    chromosome 1, pos 9
```

Burroughs-Wheeler Aligner



<table>
<tr><th colspan="5">Transformation</th></tr>
<tr><th>Input</th><th>All
Rotations</th><th>Sorting All Rows in Alphabetical
Order by their first letters</th><th>Taking
Last Column</th><th>Output
Last Column</th></tr>
<tr><td>^BANANA|</td><td>^BANANA|
|^BANANA
A|^BANAN
NA|^BANA
ANA|^BAN
NANA|^BA
ANANA|^B
BANANA|^</td><td>ANANA|^B
ANA|^BAN
A|^BANAN
BANANA|^
NANA|^BA
NA|^BANA
^BANANA|
|^BANANA</td><td>ANANA|^B
ANA|^BAN
A|^BANAN
BANANA|^
NANA|^BA
NA|^BANA
^BANANA|
|^BANANA</td><td>BNN^AA|A</td></tr>
</table>

algorithm used in computer science for file compression
original sequence can be reconstructed

BWA (`module add bwa`) Burroughs-Wheeler Aligner

Input to mapping

Reference genome

Reference.fasta

Reference.fai

```
>Potra000002
CACGAGGTTTCATCATGGACTTGGCACCATAAAA
GTTCTCTTTCATTATATTCCCTTTAGGTAAAATG
ATTCTCGTTCATTTGATAATTTTGTAATAACCGG
CCTCATTCAACCCATGATCCGACTTGATGGTGAA
TACTTGTGTAATAACTGATAATTTACTGTGATTT
ATATAACTATCTCATAATGGTTCGTCAAAATCTT
TTAAAAGATAAAAAAAACCTTTATCAATTATCTA
TATAAATTCAAATTTGTACACATTTACTAGAAAT
TACAACTCAGCAATAAAATTGACAAAATATAAAA
CAGAACCGTTAAATAAGCTATTATTTATTTCATC
ACAAAACATCTAAGTCAAAAATTTGACATAAGTT
TCATCAATTTACAAACAAACACAATTTTACAAAA
TCTCAACCAAACCATAACATGTACAAATTATAAA
TATCAACAATATTGTTTGAGAAAAAAACTATAAC
ACAAGTAAATACCAAAAAAAATACATATACTACA
AAACAATATATAAAAAATTAACATTTTAAAATTG
TGTTCAAATAAAAAATTAGATTTGCTTACTTAAG
CTGGAGAATTCTGAATAAAATTTGAATTAGAAGA
```

Sample data

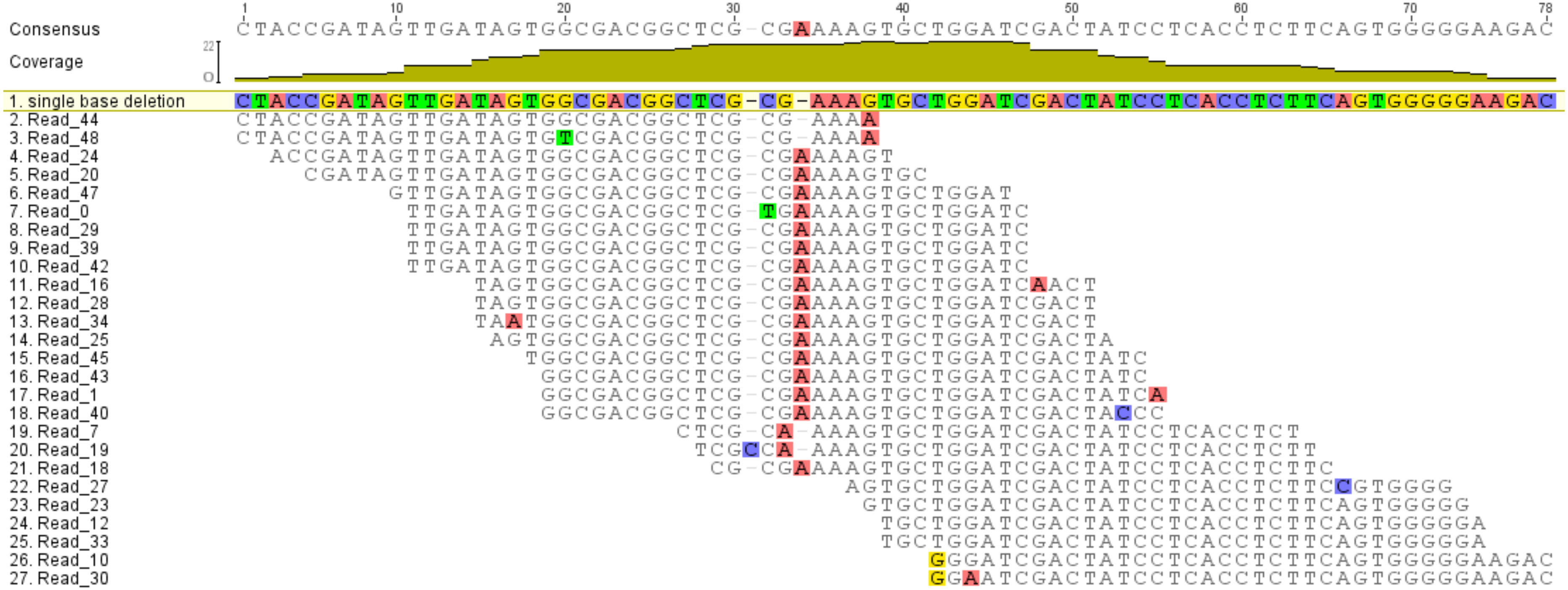
R1.fastq

R2.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFHHHHHGJJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
@HISEQ:100:C3MG8ACXX:
5:1101:1448:2164 1:N:0:ATCACG
NAGATTGTTTGTGTGCCTAAATAAATAAATAAAT
AAAAATGATGATGGTCTTAAAGGAATTTGAAATT
AAGATTGAGATATTGAAAAAGCAGATGTGGTC
+
#1=DDFFEHHDFHHJGGIJJJJGIHIGIJJJJJI
IJJJJIJJJFIJJF?
FHHHIIJJIIJJIGIIJJJIJIGHGHIIJJIHGH
GHGHFFFFDFFF>CDDD
```

Output from mapping

SciLifeLab



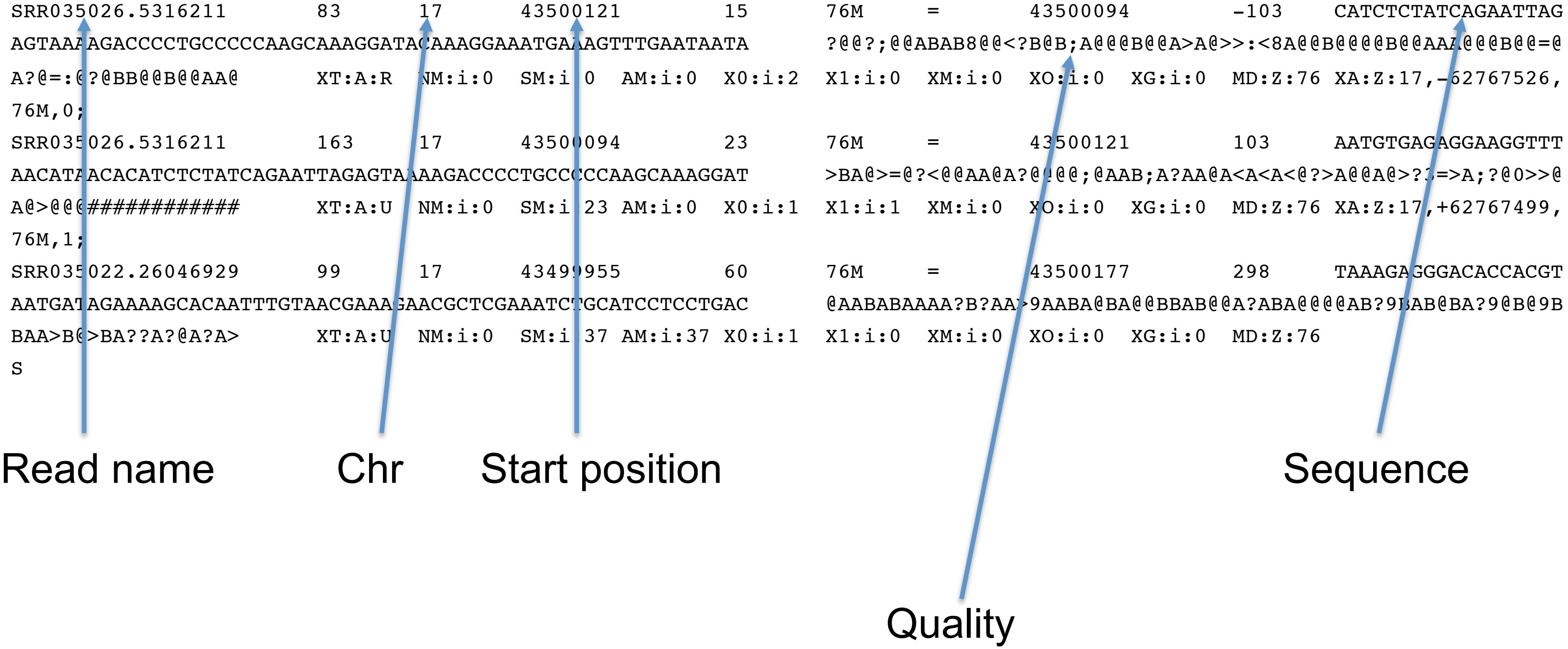
```
                          1        10        20        30        40        50        60        70      78
Consensus                 CTACCGATAGTTGATAGTGGCGACGGCTCG-CGAAAAGTGCTGGATCGACTATCCTCACCTCTTCAGTGGGGGAAGAC
Coverage              22
                       0
1. single base deletion   CTACCGATAGTTGATAGTGGCGACGGCTCG-CG-AAAGTGCTGGATCGACTATCCTCACCTCTTCAGTGGGGGAAGAC
2. Read_44                CTACCGATAGTTGATAGTGGCGACGGCTCG-CG-AAAA
3. Read_48                CTACCGATAGTTGATAGTGTCGACGGCTCG-CG-AAAA
4. Read_24                  ACCGATAGTTGATAGTGGCGACGGCTCG-CGAAAAGT
5. Read_20                    CGATAGTTGATAGTGGCGACGGCTCG-CGAAAAGTGC
6. Read_47                         GTTGATAGTGGCGACGGCTCG-CGAAAAGTGCTGGAT
7. Read_0                           TTGATAGTGGCGACGGCTCG-TGAAAAGTGCTGGATC
8. Read_29                          TTGATAGTGGCGACGGCTCG-CGAAAAGTGCTGGATC
9. Read_39                          TTGATAGTGGCGACGGCTCG-CGAAAAGTGCTGGATC
10. Read_42                         TTGATAGTGGCGACGGCTCG-CGAAAAGTGCTGGATC
11. Read_16                             TAGTGGCGACGGCTCG-CGAAAAGTGCTGGATCAACT
12. Read_28                             TAGTGGCGACGGCTCG-CGAAAAGTGCTGGATCGACT
13. Read_34                             TAATGGCGACGGCTCG-CGAAAAGTGCTGGATCGACT
14. Read_25                              AGTGGCGACGGCTCG-CGAAAAGTGCTGGATCGACTA
15. Read_45                                TGGCGACGGCTCG-CGAAAAGTGCTGGATCGACTATC
16. Read_43                                 GGCGACGGCTCG-CGAAAAGTGCTGGATCGACTATC
17. Read_1                                  GGCGACGGCTCG-CGAAAAGTGCTGGATCGACTATCA
18. Read_40                                 GGCGACGGCTCG-CGAAAAGTGCTGGATCGACTACCC
19. Read_7                                          CTCG-CA-AAAGTGCTGGATCGACTATCCTCACCTCT
20. Read_19                                          TCGCCA-AAAGTGCTGGATCGACTATCCTCACCTCTT
21. Read_18                                           CG-CGAAAAGTGCTGGATCGACTATCCTCACCTCTTC
22. Read_27                                                   AGTGCTGGATCGACTATCCTCACCTCTTCCGTGGGG
23. Read_23                                                    GTGCTGGATCGACTATCCTCACCTCTTCAGTGGGGG
24. Read_12                                                     TGCTGGATCGACTATCCTCACCTCTTCAGTGGGGGA
25. Read_33                                                     TGCTGGATCGACTATCCTCACCTCTTCAGTGGGGGA
26. Read_10                                                        GGGATCGACTATCCTCACCTCTTCAGTGGGGGAAGAC
27. Read_30                                                        GGAATCGACTATCCTCACCTCTTCAGTGGGGGAAGAC
```

Output - SAM format

HEADER SECTION

```
@SQ     SN:17   LN:81195210
@PG     ID:bwa  PN:bwa  VN:0.7.13-r1126 CL:bwa sampe human_17_v37.fasta NA06984.ILLUMINA.low_coverage.17_1.sai NA06984.ILL
UMINA.low_coverage.17_2.sai /proj/g2016008/labs/gatk/fastq/wgs/NA06984.ILLUMINA.low_coverage.17q_1.fq /proj/g2016008/labs/
gatk/fastq/wgs/NA06984.ILLUMINA.low_coverage.17q_2.fq
```

ALIGNMENT SECTION



```
SRR035026.5316211       83      17      43500121        15      76M     =       43500094        -103    CATCTCTATCAGAATTAG
AGTAAAAGACCCCTGCCCCCAAGCAAAGGATACAAAGGAAATGAAAGTTTGAATAATA      ?@@?;@@ABAB8@@<?B@B;A@@@B@@A>A@>>:<8A@@B@@@@B@@AAA@@@B@@=@
A?@=:@?@BB@@B@@AA@      XT:A:R  NM:i:0  SM:i:0  AM:i:0  X0:i:2  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76 XA:Z:17,-62767526,
76M,0;
SRR035026.5316211       163     17      43500094        23      76M     =       43500121        103     AATGTGAGAGGAAGGTTT
AACATAACACATCTCTATCAGAATTAGAGTAAAAGACCCCTGCCCCCAAGCAAAGGAT      >BA@>=@?<@@AA@A?@@@@;@AAB;A?AA@A<A<A<@?>A@@A@>?3=>A;?@0>>@
A@>@@@############      XT:A:U  NM:i:0  SM:i:23 AM:i:0  X0:i:1  X1:i:1  XM:i:0  XO:i:0  XG:i:0  MD:Z:76 XA:Z:17,+62767499,
76M,1;
SRR035022.26046929      99      17      43499955        60      76M     =       43500177        298     TAAAGAGGGACACCACGT
AATGATAGAAAAGCACAATTTGTAACGAAAGAACGCTCGAAATCTGCATCCTCCTGAC      @AABABAAAA?B?AA>9AABA@BA@@BBAB@@A?ABA@@@@AB?9BAB@BA?9@B@9B
BAA>B@>BA??A?@A?A>      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:76
S
```

Read name

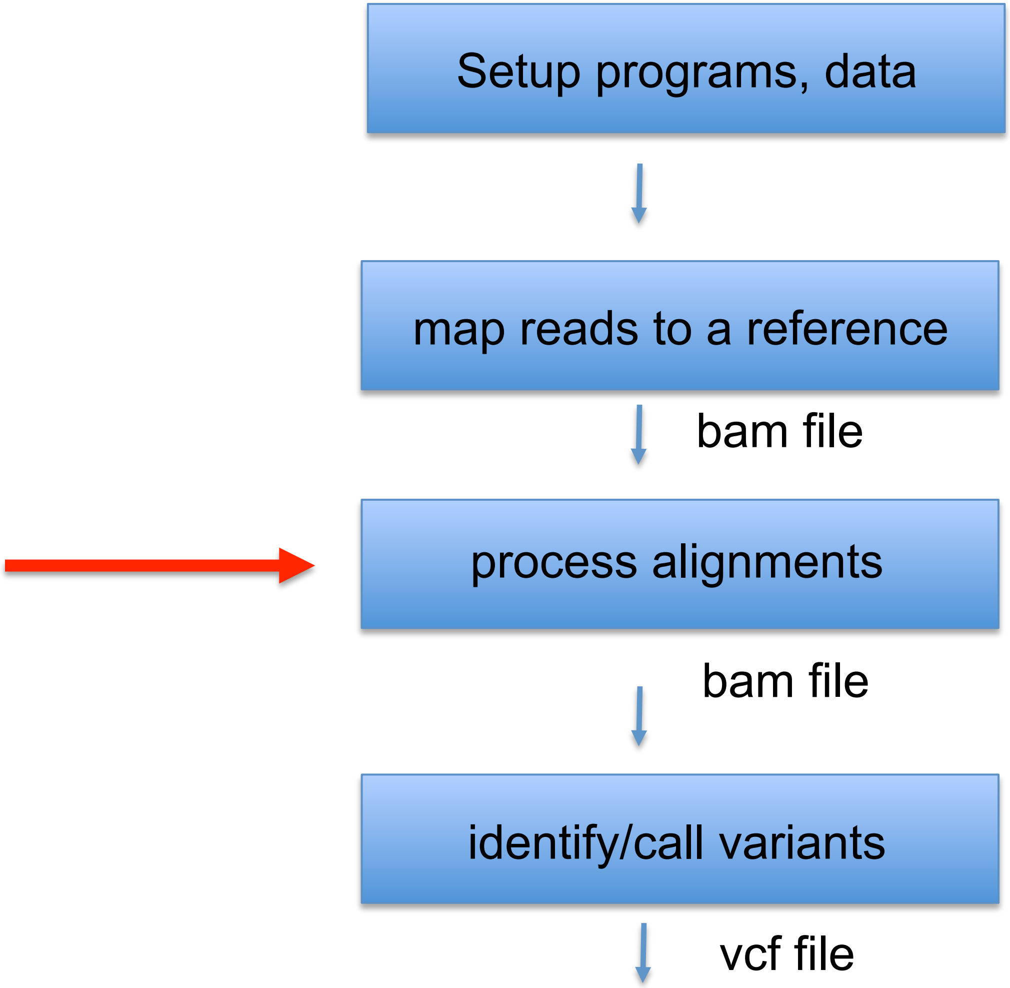
Chr

Start position

Sequence

Quality

Steps in resequencing analysis



Setup programs, data

map reads to a reference

find best placement of reads

bam file

process alignments

realign indels
remove duplicates
recalibrate base quality

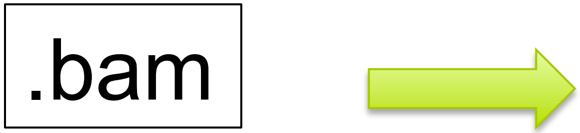
bam file

identify/call variants

statistical algorithms
to detect true variants

vcf file

Processing BAM files

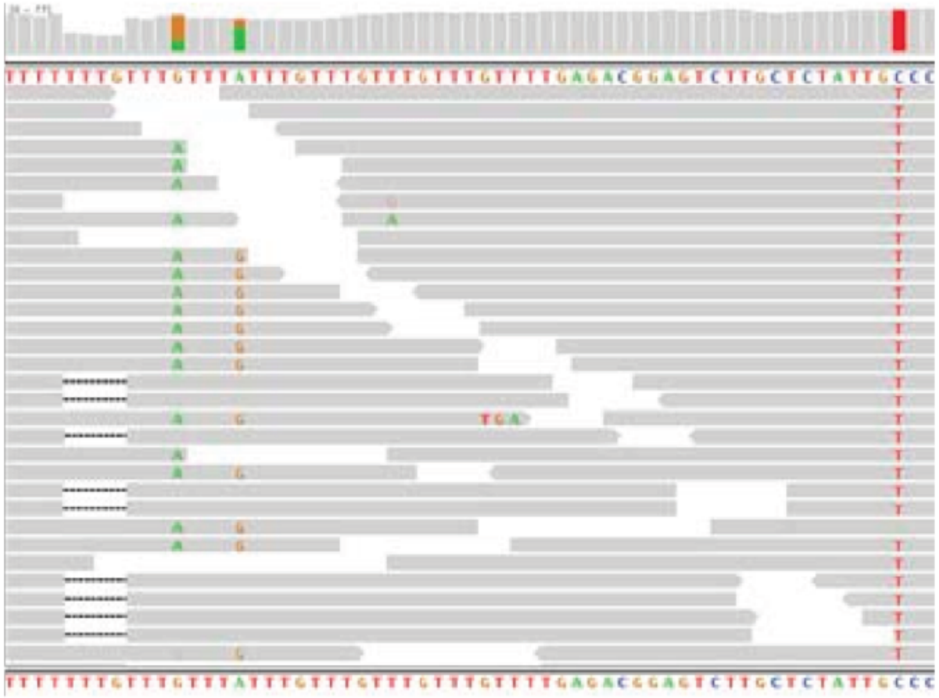


.bam

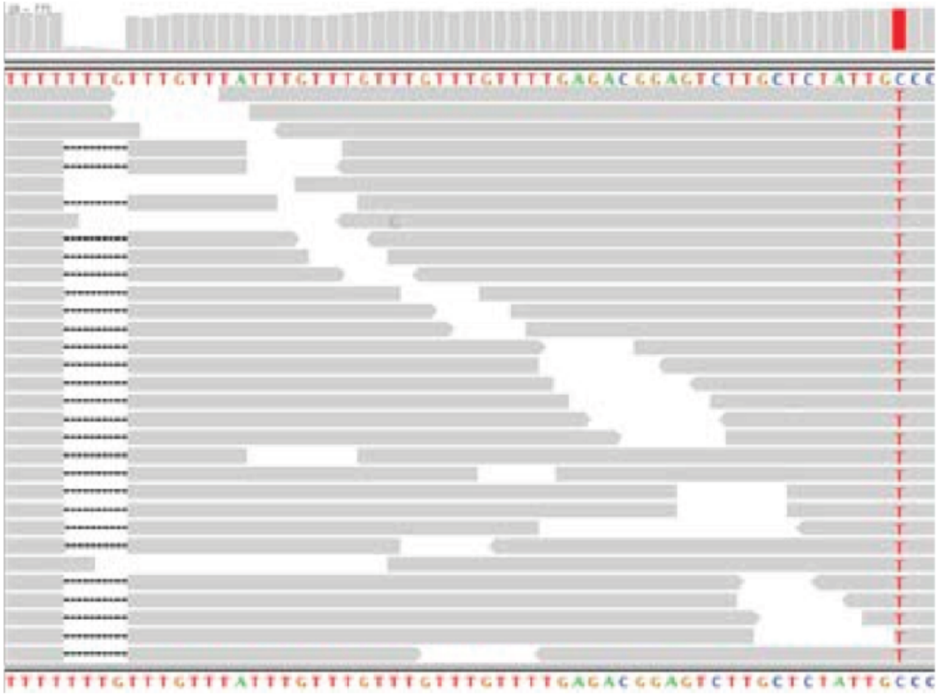
Realign around indels

Realign around indels

- mapping is done one read at a time
- single variants may be split into multiple variants
- solution: realign these regions taking all reads into account

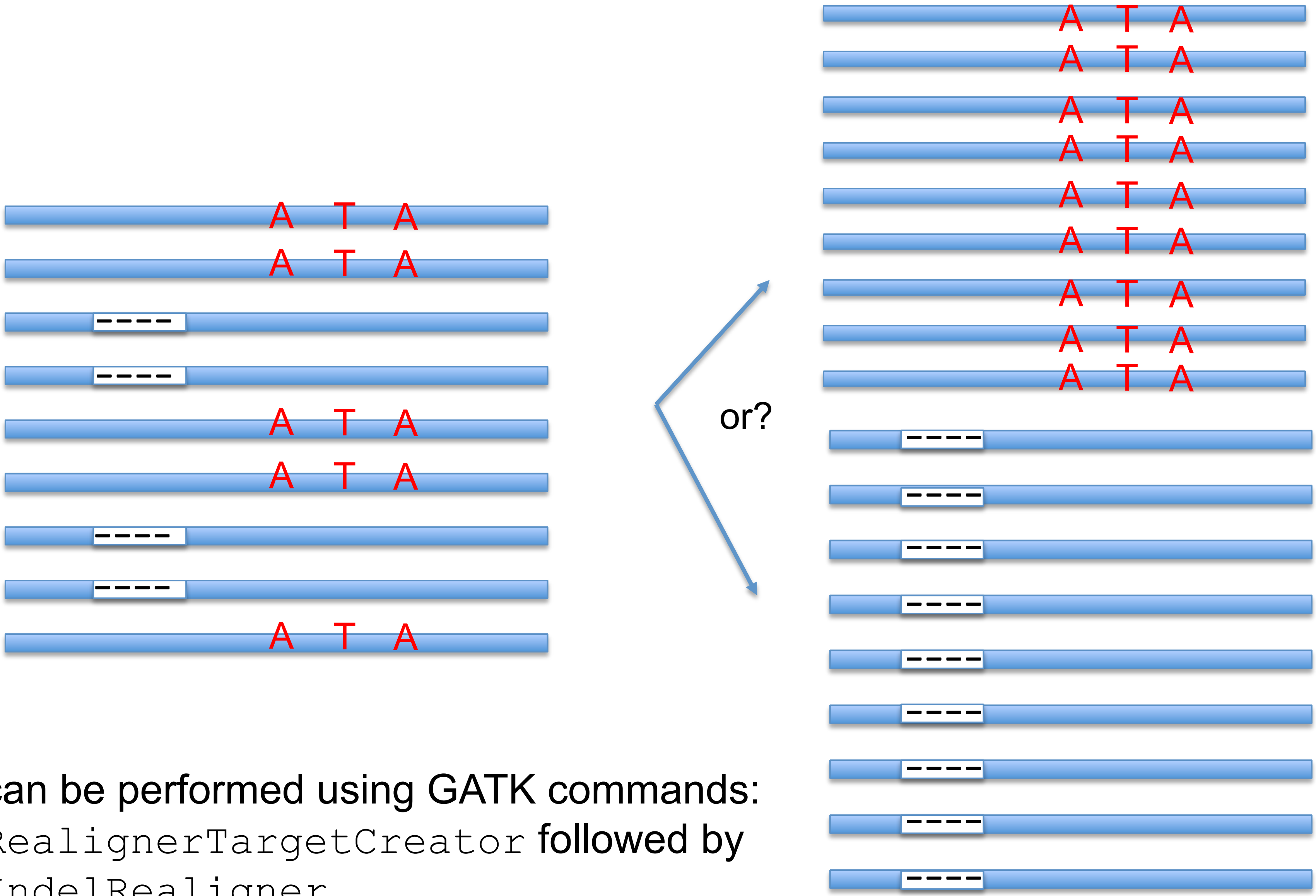

TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC

HiSeq data, raw BWA alignments

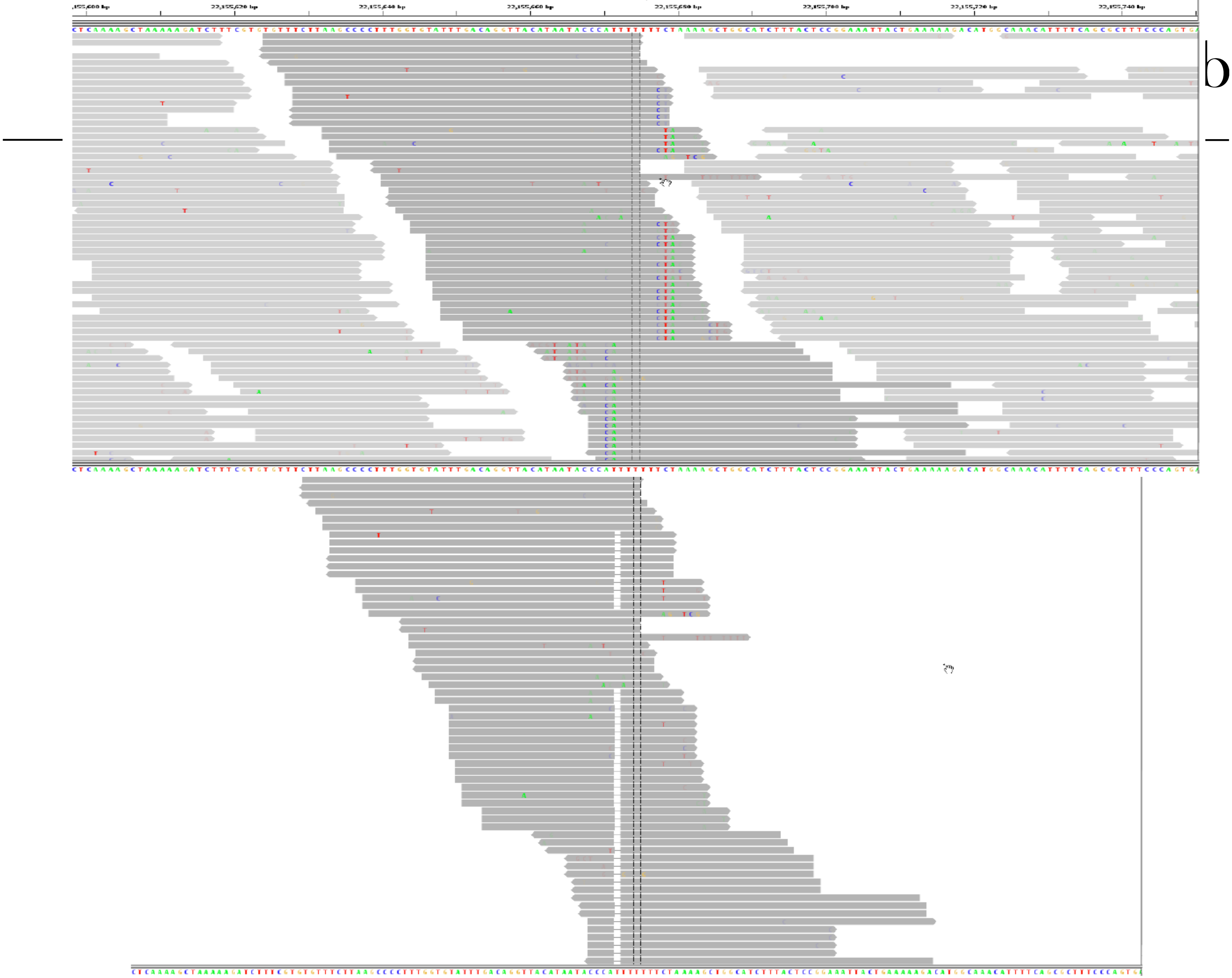

TTTTTTTGTTTGTTTATTTGTTTGTTTGTTTGTTTTGAGACGGAGTCTTGCTCTATTGCCC

HiSeq data, after MSA

Local realignment



can be performed using GATK commands: `RealignerTargetCreator` followed by `IndelRealigner`


b
,155,600 bp
22,155,620 bp
22,155,640 bp
22,155,660 bp
22,155,680 bp
22,155,700 bp
22,155,720 bp
22,155,740 bp
CTCAAAAGCTAAAAAGATCTTTCGTGTGTTTCTTAAGCCCCTTTGGTGTATTTGACAGGTTACATAATACCCATTTTTTTTCTAAAAGCTGGCATCTTTACTCCGGAAATTACTGAAAAAGACATGGCAAACATTTTCAGCGCTTTCCCAGTGA

Remove duplicates

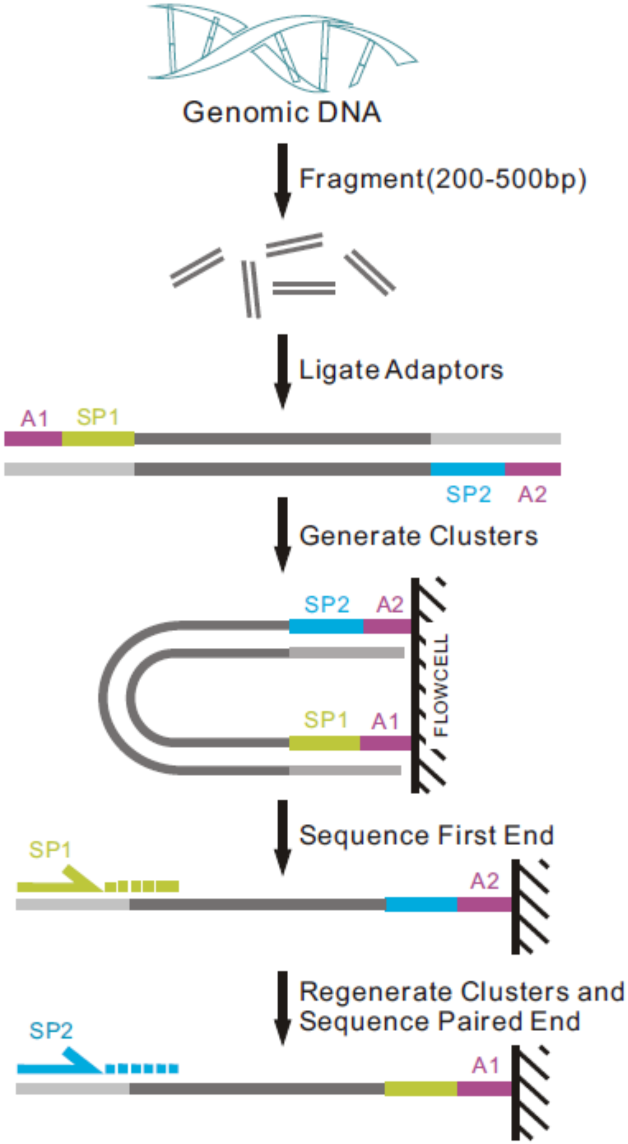
PCR duplicates

- The same DNA fragment sequenced multiple times
  - not independent observations
  - skew allele frequency and read depth
  - errors double counted

- PCR duplicates occur
  - during library prep, or
  - optical duplicates (one cluster read as two)

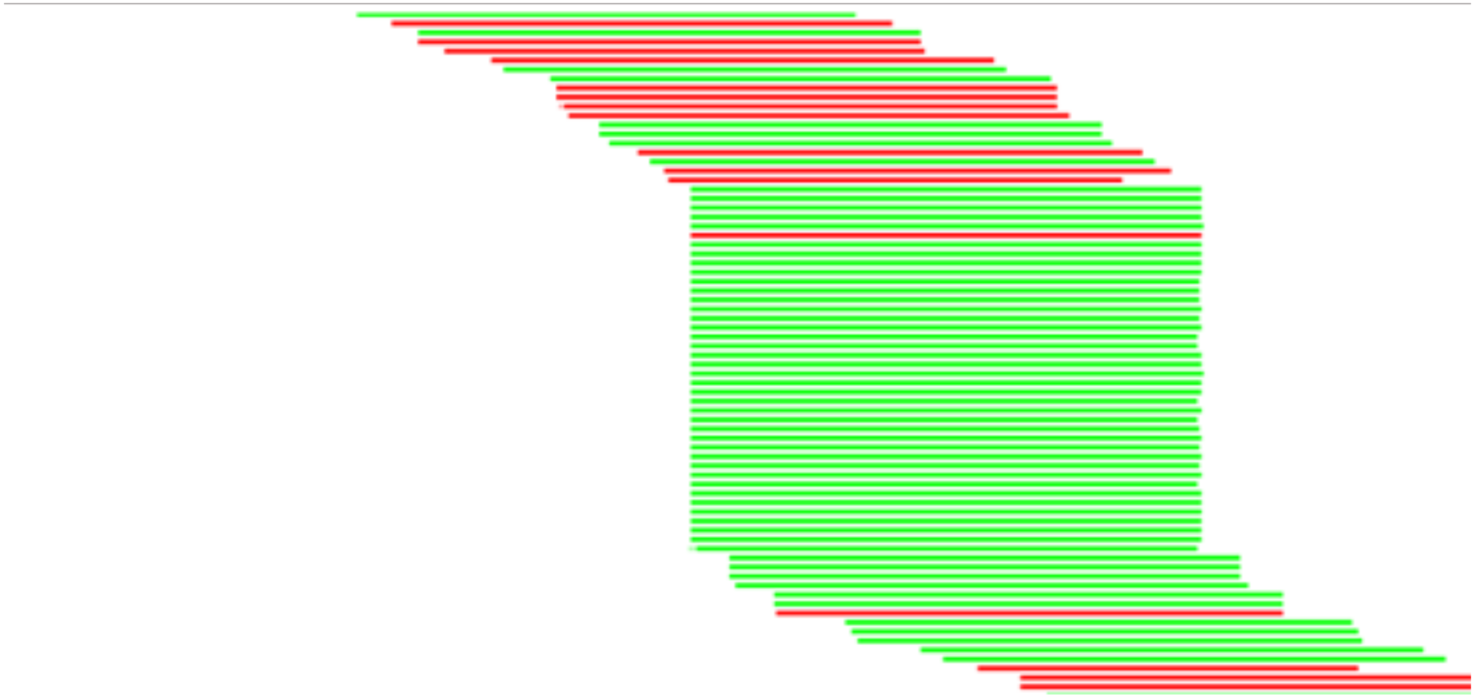
- Reading: http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Paired end sequencing



Identify PCR duplicates

- Single or paired reads that map to identical positions
- Mark and/or remove them!
- Picard `MarkDuplicates`



Base quality score recalibration

Base quality scores are per-base estimates of error emitted by the sequencing machines (i.e. probability that the called base is wrong).

Scores produced by the machines are subject to various sources of systematic technical error, leading to over- or under-estimated base quality scores in the data.

Base quality score recalibration

1. Empirically models errors in the quality scores using a machine learning process
2. Adjusts the quality scores to minimize errors

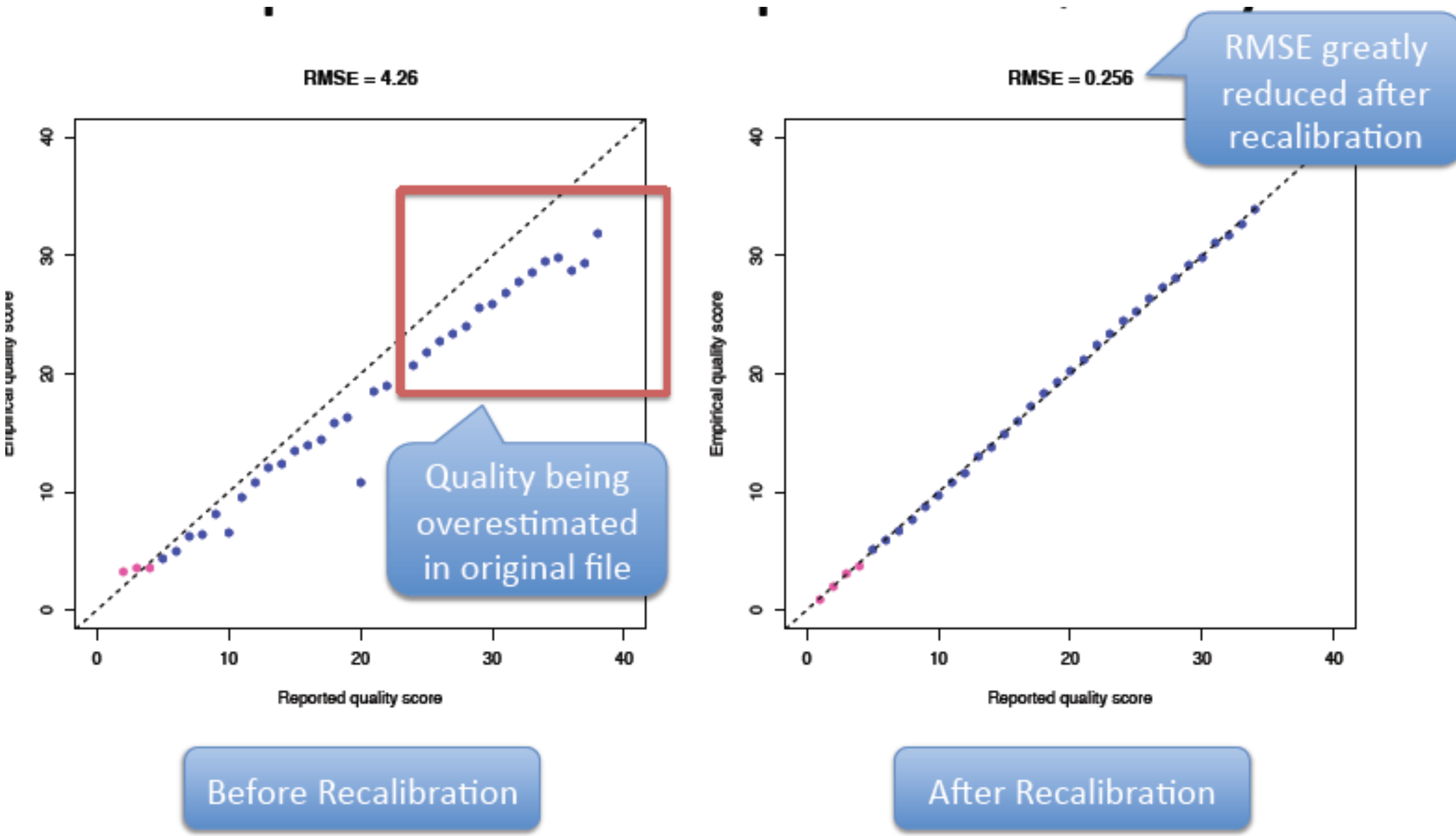
Empirical modeling of error in quality score
At a given position in the genome:
Compare $RMSE = (Qualityscore - EmpiricalScore)2$
The average base quality scores over all reads
With
Observed error rate, i.e. fraction of reads that differ from the reference genome sequence **at non-polymorphic sites**

RMSE = Root mean square error
Measure of the difference between predicted values and the values actually observed
i.e. base qualities vs fraction of reads that differ from reference

Base quality score recalibration

After recalibration, the quality scores in the QUAL field in the output BAM are more accurate in that the reported quality score is closer to its actual probability of mismatching the reference genome.

Results from BQSR



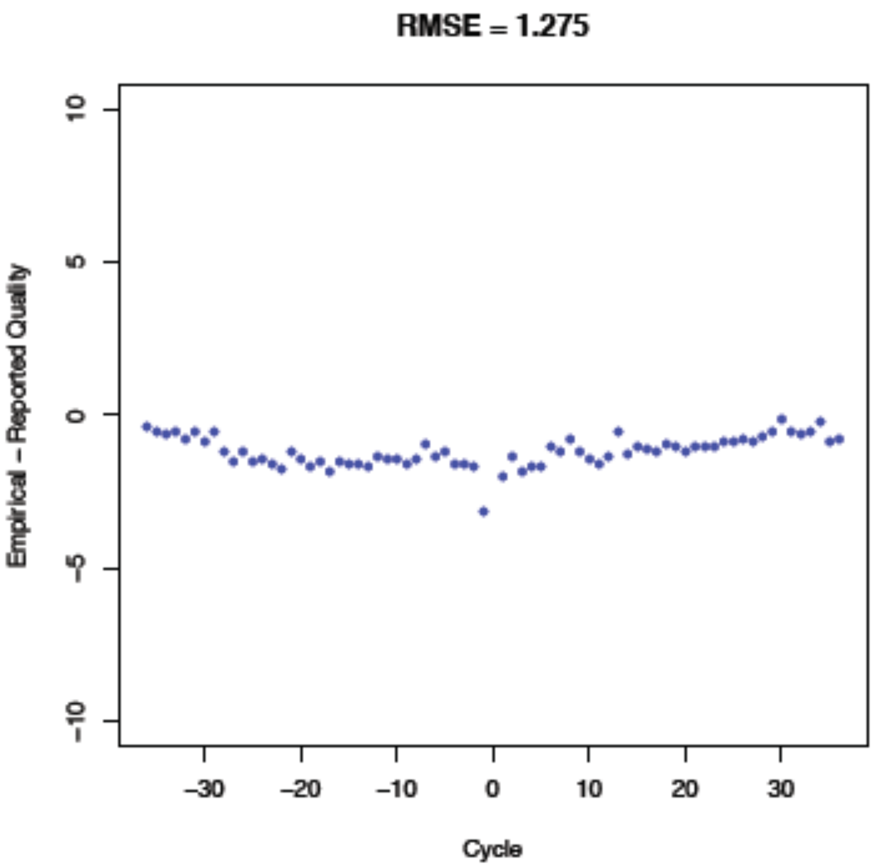
Quality being overestimated in original file

RMSE greatly reduced after recalibration

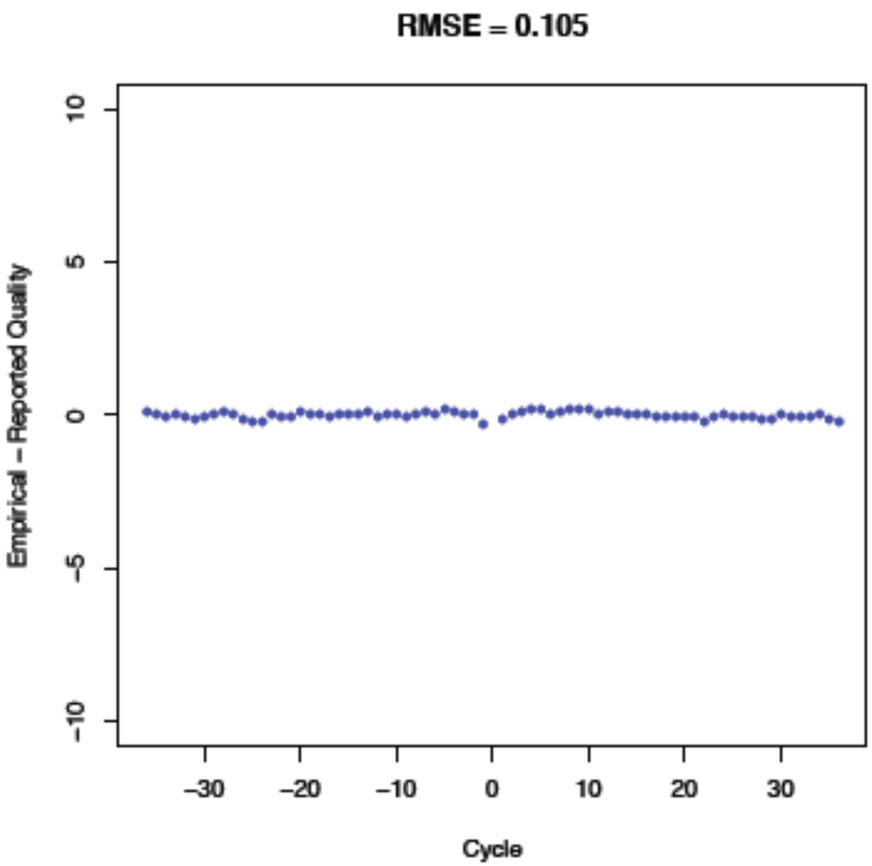
Before Recalibration

After Recalibration

Residual error by machine cycle

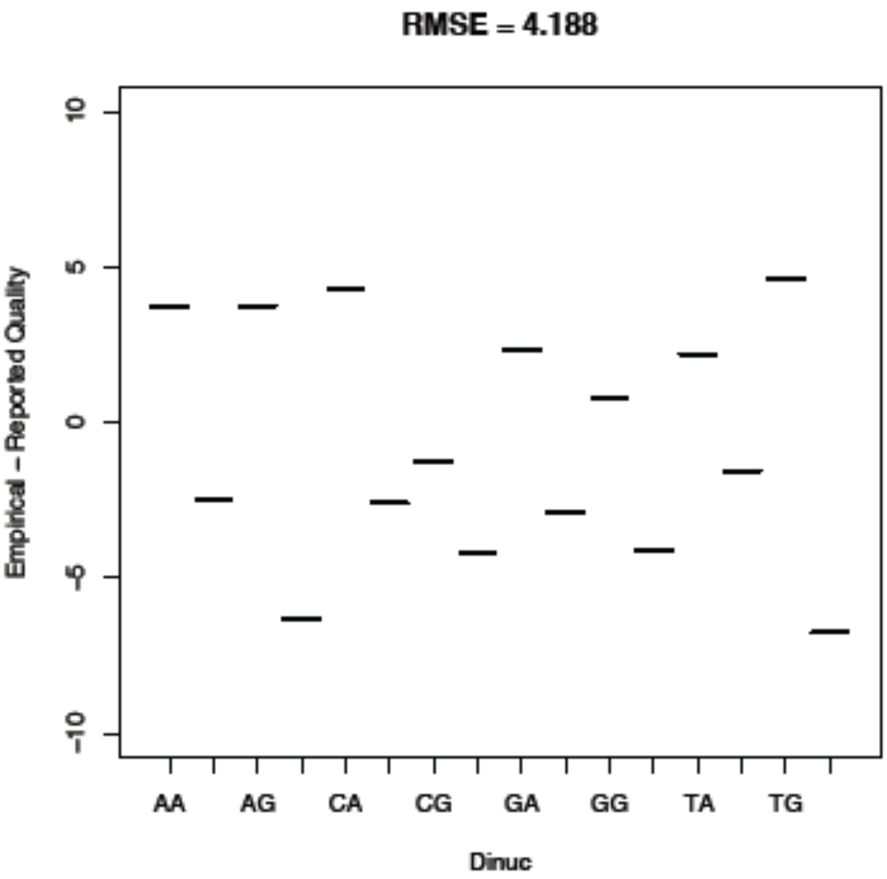


Before Recalibration

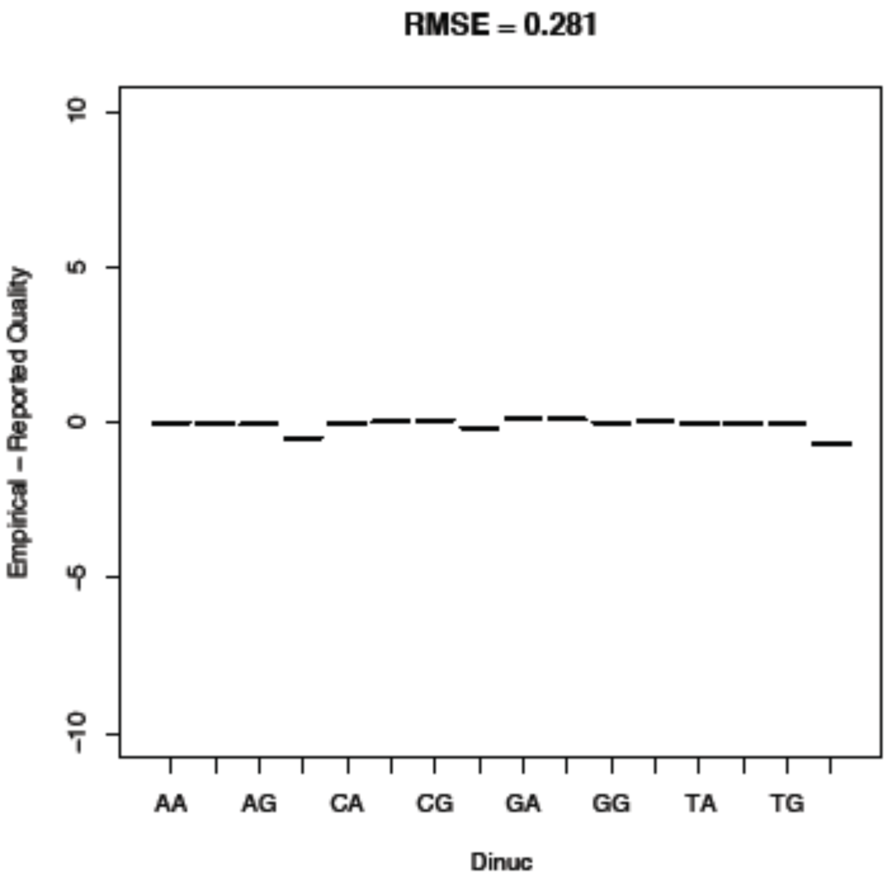


After Recalibration

Residual error by dinucleotide

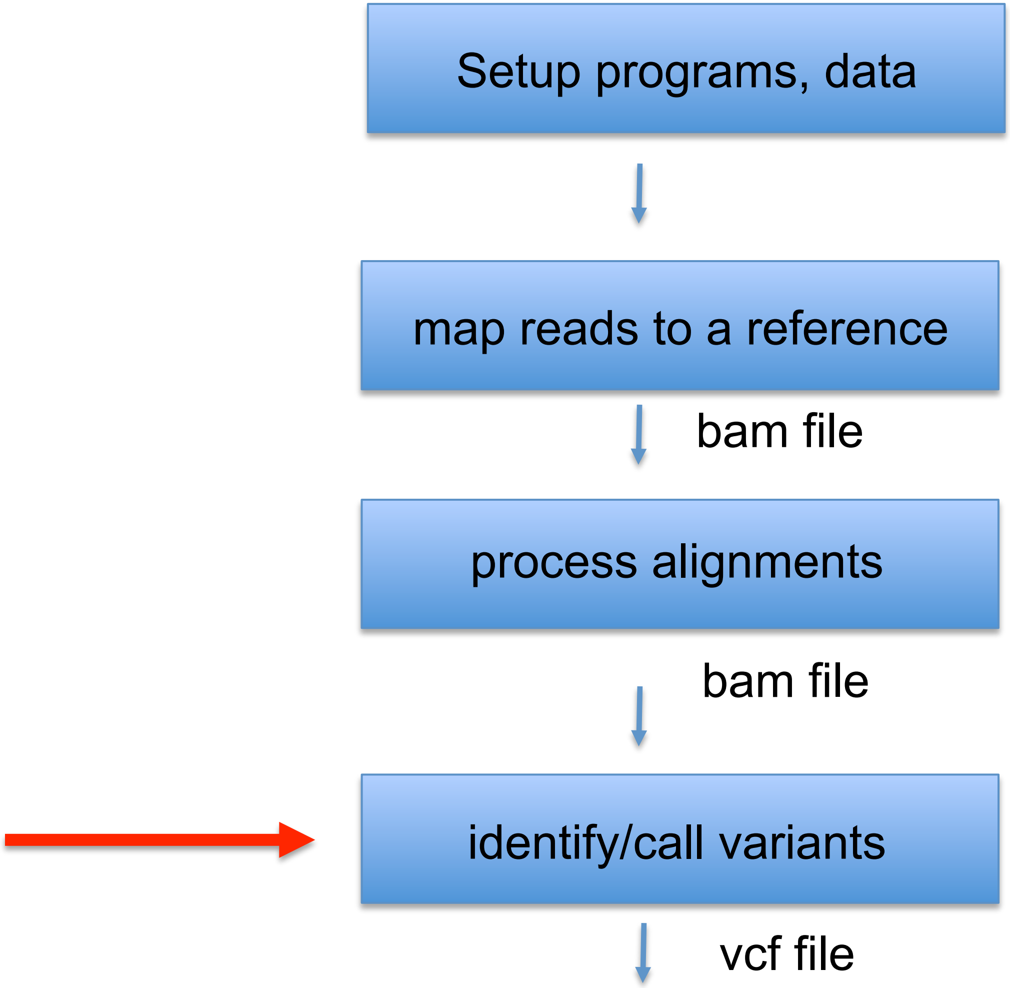


Before Recalibration



After Recalibration

Steps in resequencing analysis



Setup programs, data

map reads to a reference

bam file

process alignments

bam file

identify/call variants

vcf file

find best placement of reads

realign indels
remove duplicates
recalibrate base quality

statistical algorithms
to detect true variants

Variant calling

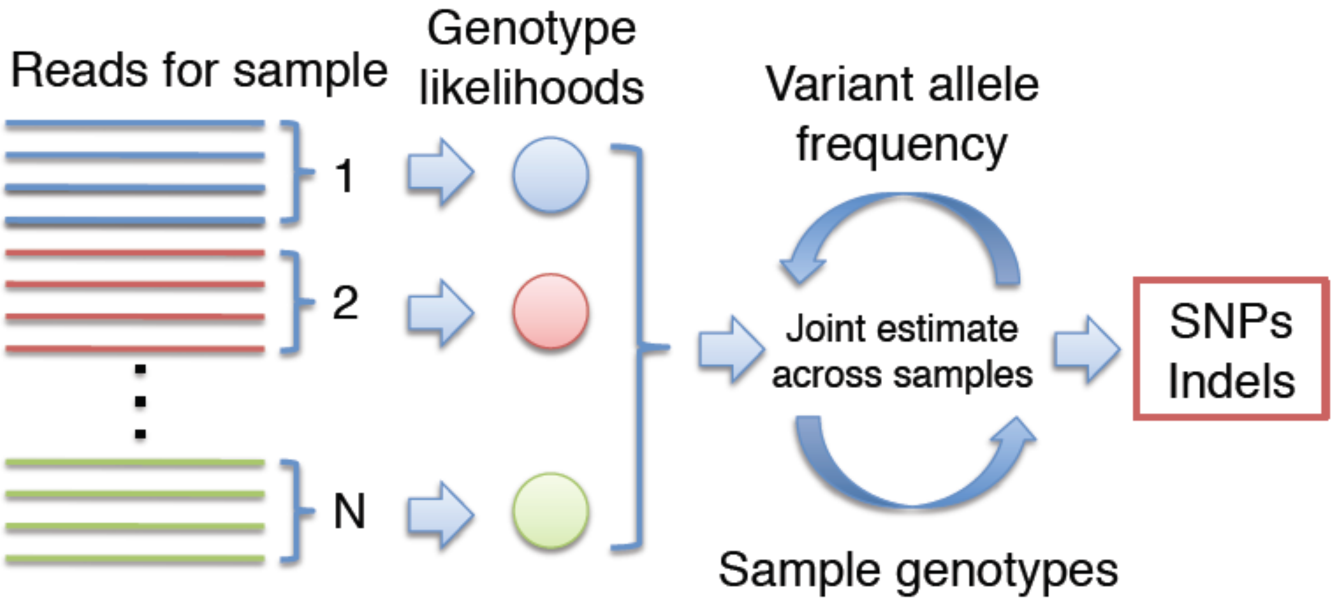
simple pileup methods

SciLifeLab

```
Reference:  acacagatagacatagacatagacagatgag

            acacagatagacatagacatagacagatgag
            acacacatagacatagacatagacagatgag
            acacagatagacatagacatagacagatgag
            acacagatagacatatacatagacagatgag
            acacagatagacatatacatagacagatgag
            acacagatagacatatacatagacagttgag
            acacagatagacatagacatagacagatgag
            acacagatagacatatacatagacagatgag
            acacagatagacatagacatagacagatgag
```

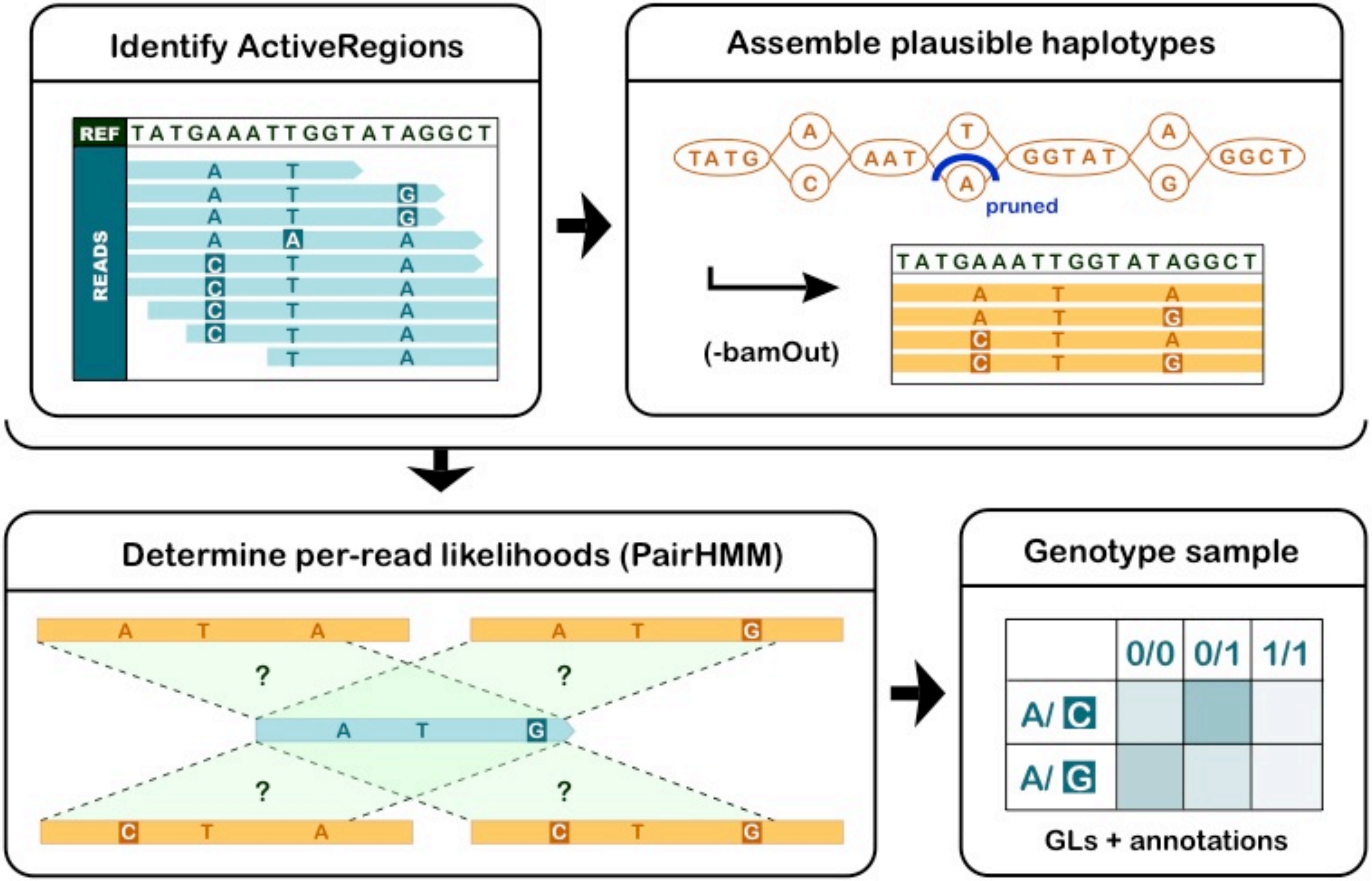
Baeysian population-based calling



Simultaneous estimation of:

- Allele frequency (AF) spectrum: Pr{AF = i I D}
- The prob. that a variant exists: Pr{AF > 0 I D}
- Assignment of genotypes to each sample

GATK haplotype caller



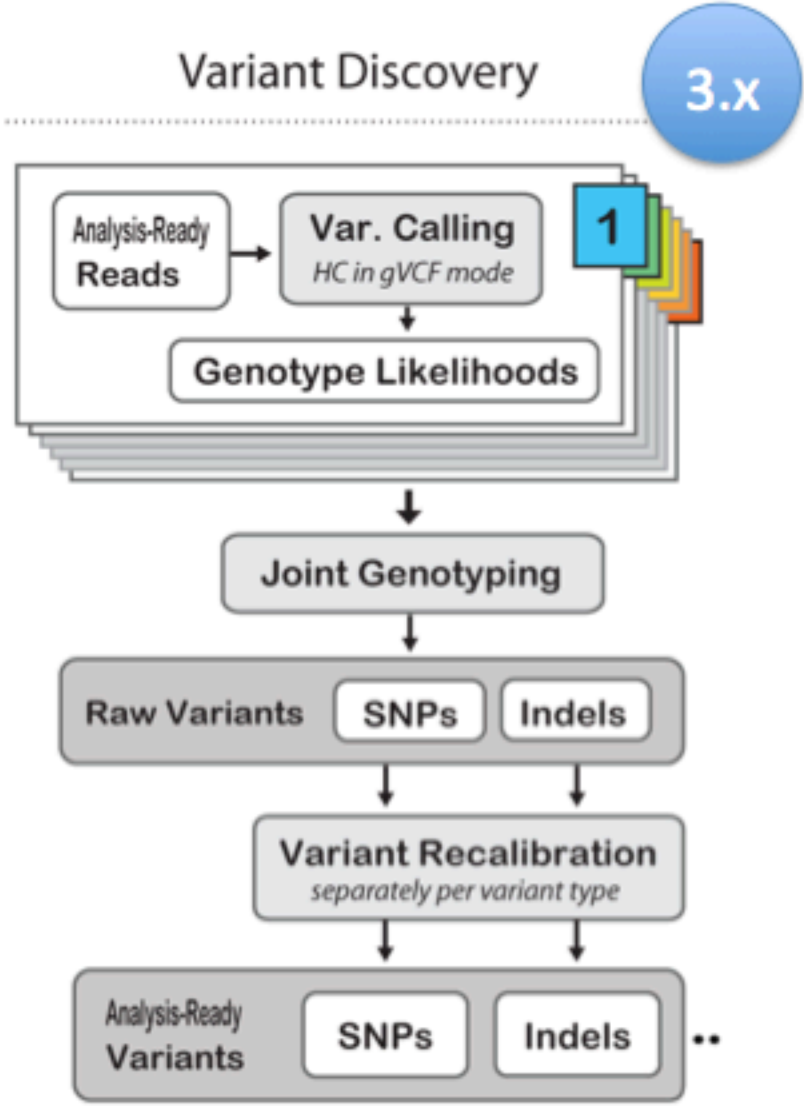
Identify ActiveRegions

Assemble plausible haplotypes

Determine per-read likelihoods (PairHMM)

Genotype sample

GATK best practice for cohorts



VCF format

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2
1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

VCF format

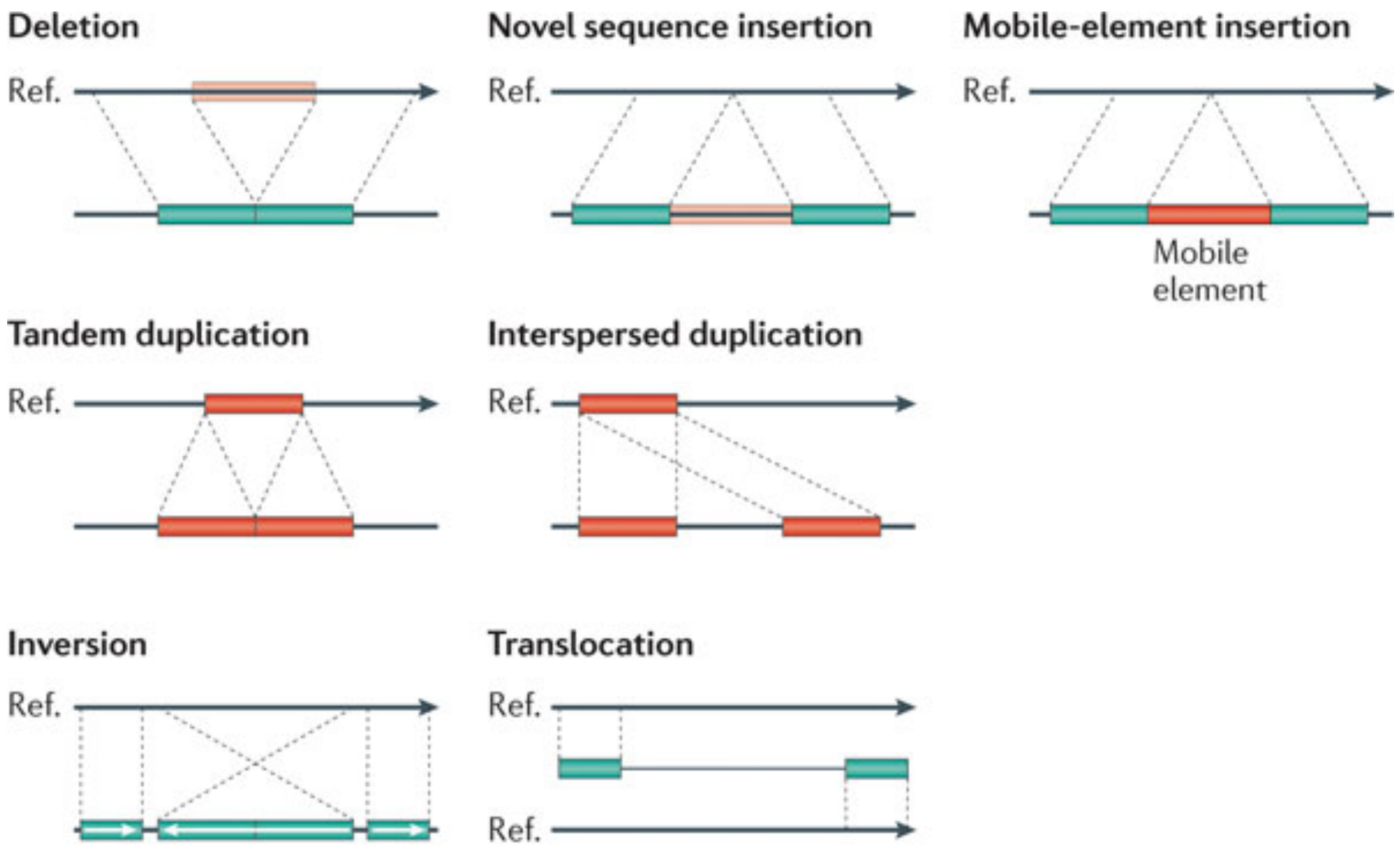
```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth"> ##INFO=<ID=AF,Number=.,Type=Float,
Frequency"> ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##FILTER=<ID=q10,Description="
##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=Intege
Quality"> ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA0
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

gVCF format

```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership"> ##FILTER=<ID=q10,Description="Quality
below 10">
##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">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
##GVCFBlock=minGQ=0(inclusive),maxGQ=5(exclusive)
##GVCFBlock=minGQ=20(inclusive),maxGQ=60(exclusive)
##GVCFBlock=minGQ=5(inclusive),maxGQ=20(exclusive)
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51
1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTCT G,GTACT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

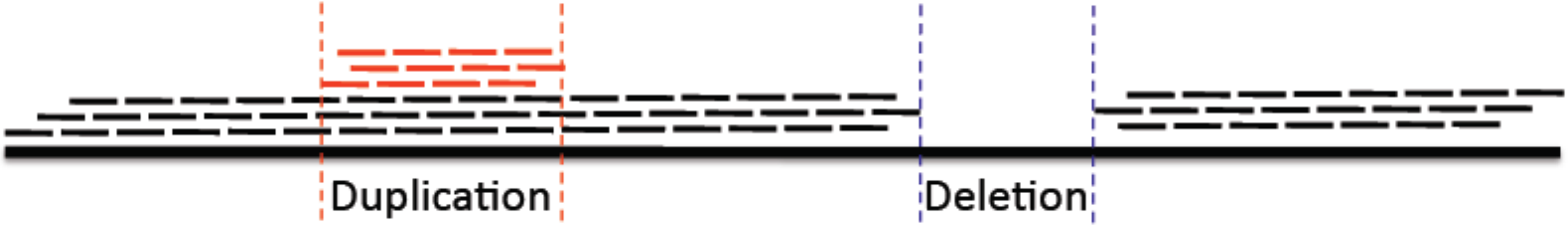
Discovery of structural variants

SciLifeLab



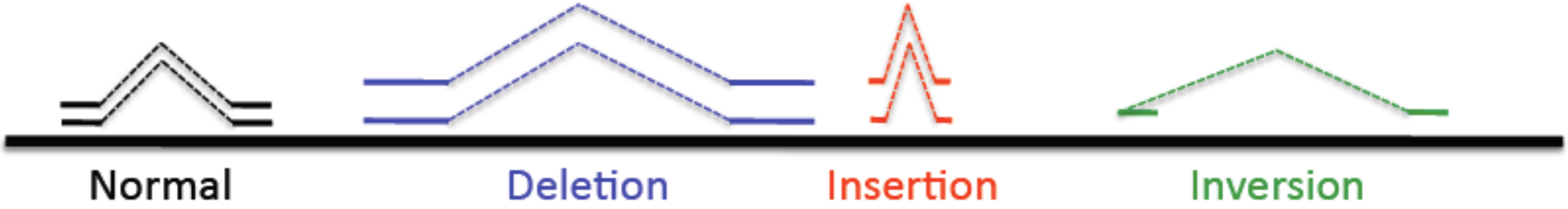
1) Read depth analysis

- Depth of coverage can be used to estimate copy number
- variation in depth indicate copy number variants
- Difficult to distinguish homozygotes and heterozygotes



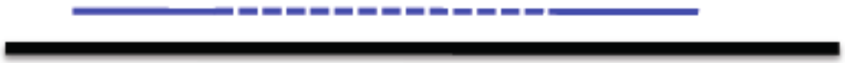
2) Paired end analysis

- Paired ends have a fixed length between them
- Genomic rearrangements cause them to vary
  - Deletion: reads will map too far apart
  - Insertion: reads will map too close
  - Inversion: reads in wrong orientation
- more reliable with long pairs

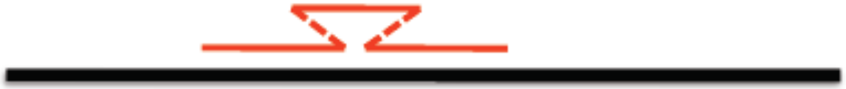


3) Split-read alignments

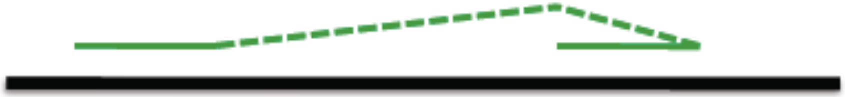
- Base-level breakpoint resolution
- Only works with long reads
  - short reads have many spurious splits
- Caveat: breakpoints may be duplicated
  - reads won't split if single alignment is good



Deletion



Insertion

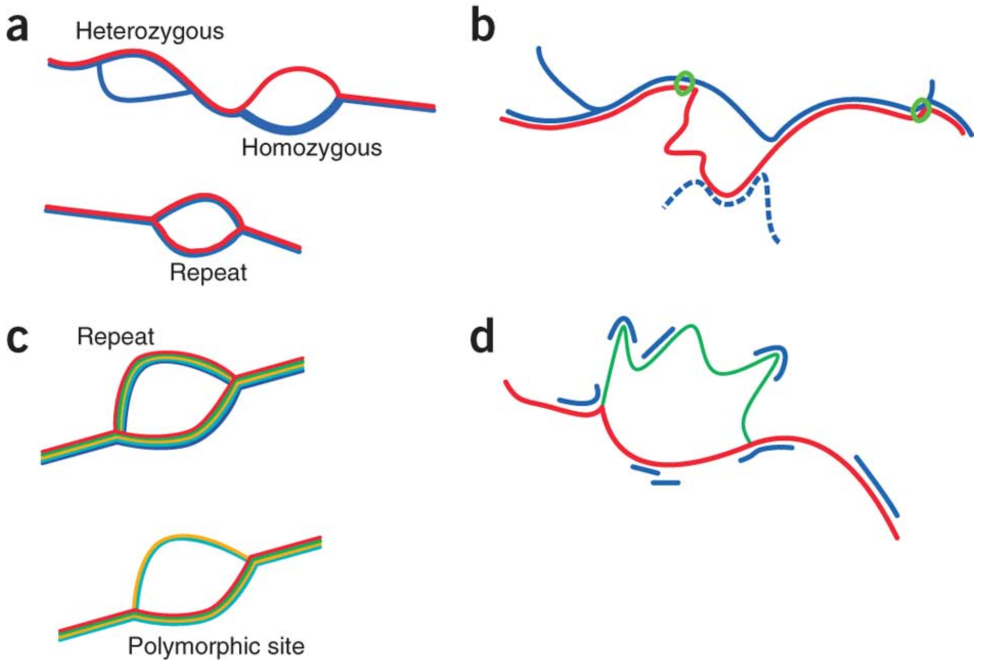


Inversion

4) *De novo* assembly to identify structural variants

SciLifeLab

- Assemble contigs
- Align to reference
- Look for insertions, deletions, rearrangements

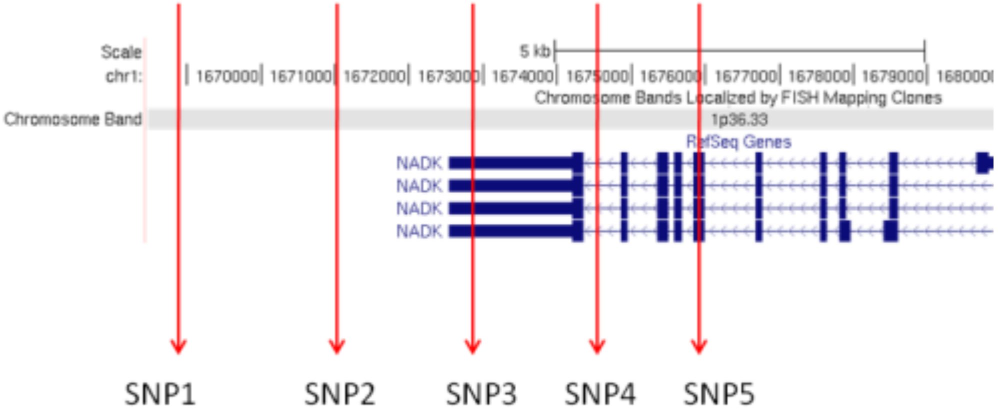


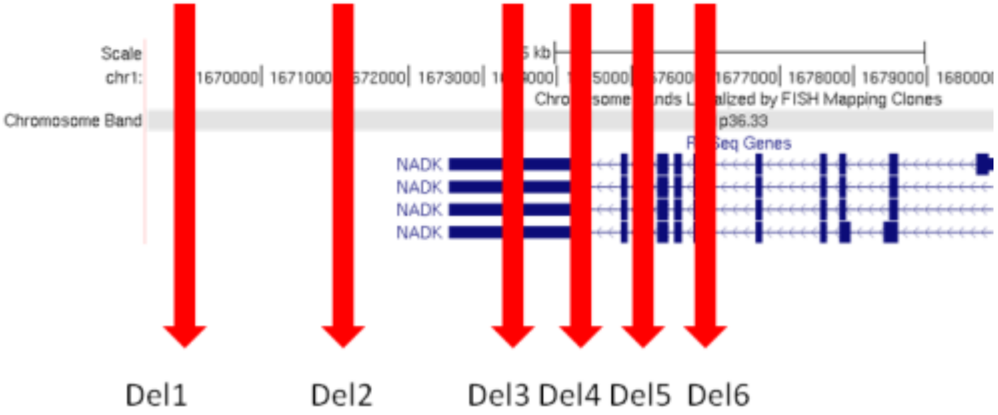
Annotation of variants

Compare variants with annotation of the reference genome

-protein coding exon
-untranslated exon
-regulatory region

Gives clues to expected effect of variant

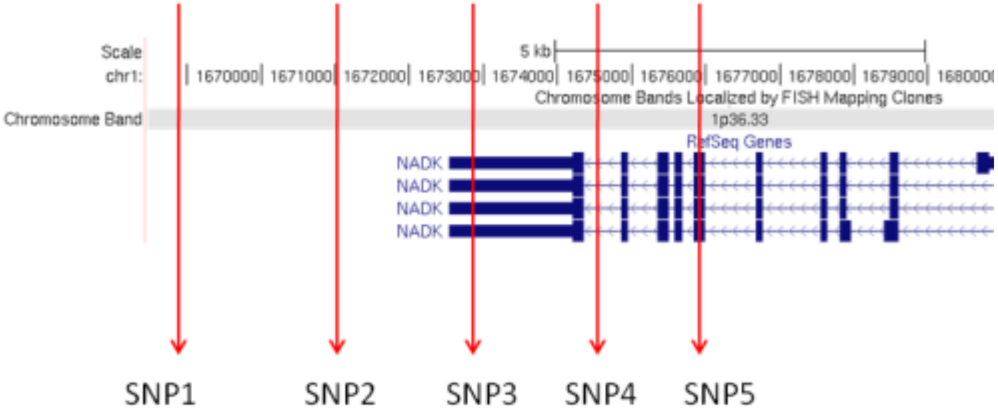




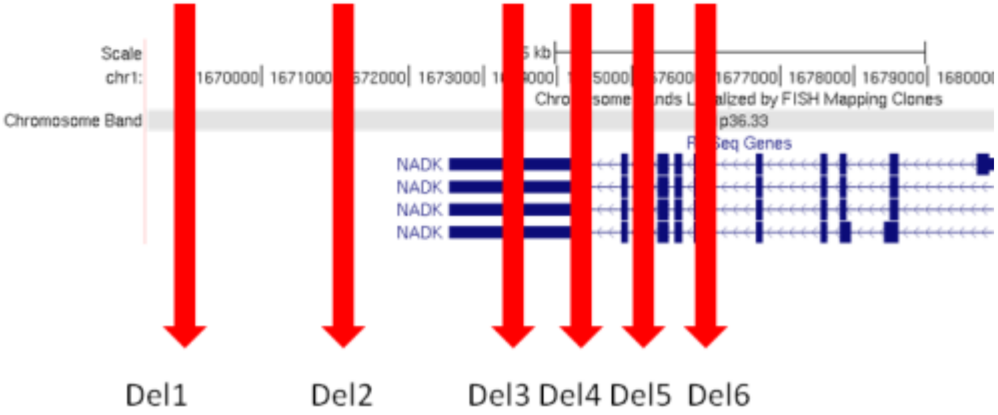
Annotation of variants

Compare variants with annotation of the reference genome

-protein coding exon
-untranslated exon
-regulatory region



Gives clues to expected effect of variant

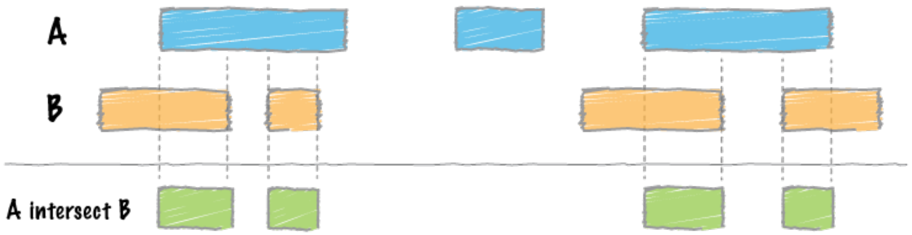


Most commonly used tools are Annovar and SNPEff

Downstream analysis

Software for file handling

- BEDTools – enables genome arithmetics – (`module add BEDTools`)

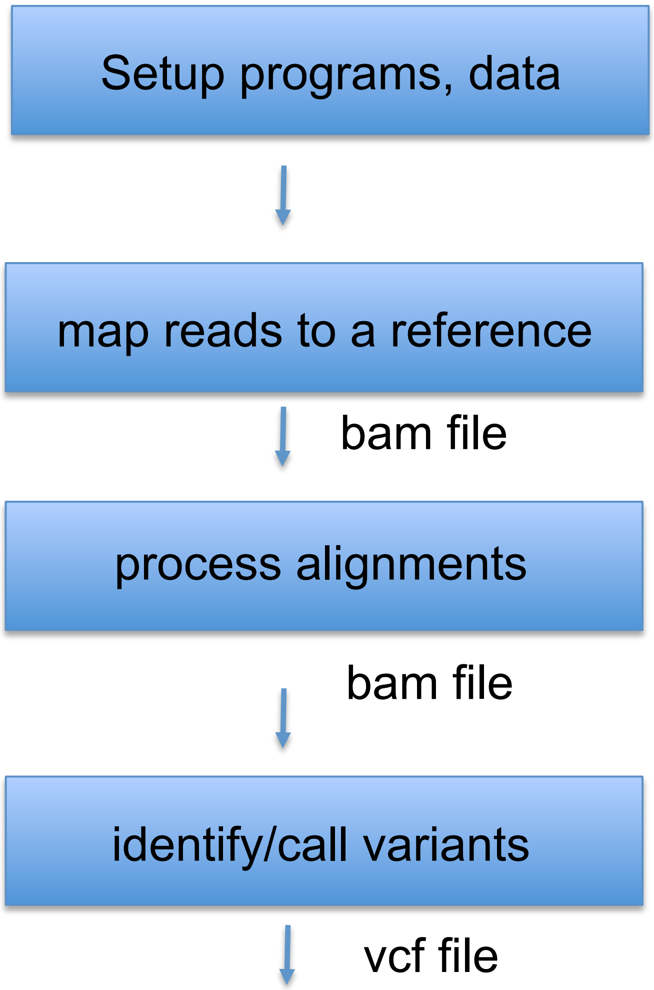


- Vcftools – for manipulations of vcf-files - (`module add vcftools`)
- bcftools – for manipulations of bcf-files - (`module add bcftools`)
- bamtools – for manipulations of bam-files - (`module add bamtools`)

Annotations to compare with can be extracted from e.g the UCSC browser, ensemble database, etc

Scripting yourself with .. Perl / python / bash / awk

Excercise



Setup programs, data

map reads to a reference

find best placement of reads

bam file

process alignments

realign indels
remove duplicates
recalibrate base quality

bam file

identify/call variants

statistical algorithms
to detect true variants

vcf file

Overview of excercise

1. Access to data and programs
2. Mapping (BWA)
3. Merging alignments (BWA)
4. Creating BAM files (Picard)
5. Processing files (GATK)
6. Variant calling and filtering (GATK)
7. Viewing data (IGV)

X. Optional extras

1) Access to data

- Data comes from 1000 genomes pilot project
  - – 81 low coverage (2-4 x) Illumina WGS samples
  - – 63 Illumina exomes
  - – 15 low coverage 454
  - – ~ 1 Mb from chromosome 17

- Fastq files located in
  - – /sw/courses/ngsintro/gatk
  - – this folder is read only

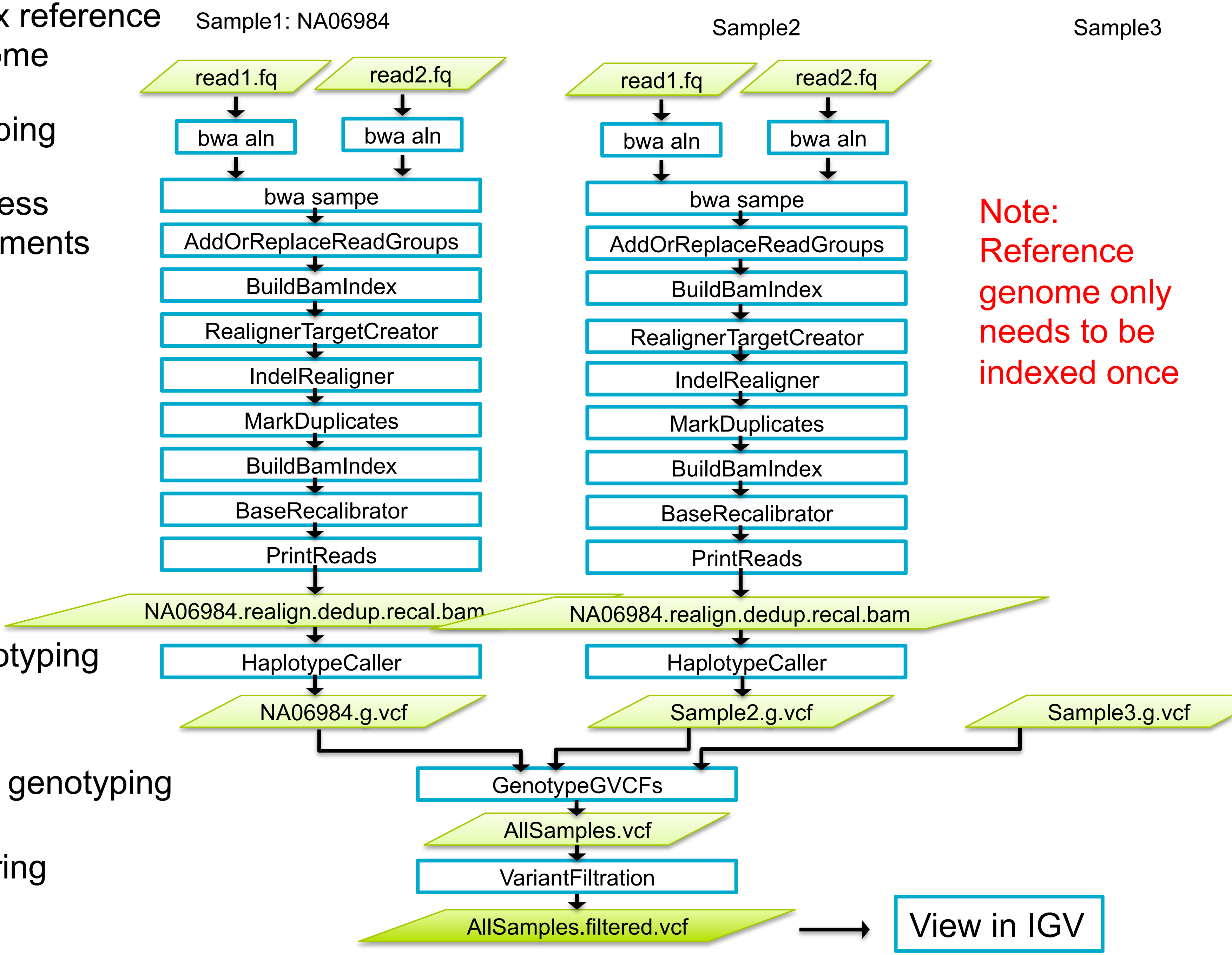
1) Access to programs

- BWA and samtools modules can be loaded:

```
module load bioinfo-tools
module load bwa
module load samtools
```

- picard and GATK are are set of java programs:

```
/bubo/sw/apps/bioinfo/GATK/3.4-46/
/bubo/sw/apps/bioinfo/picard/1.69/kalkyl/
```



Index reference genome

Sample1: NA06984

Sample2

Sample3

read1.fq

read2.fq

read1.fq

read2.fq

Mapping

bwa aln

bwa aln

bwa aln

bwa aln

Process alignments

bwa sampe

AddOrReplaceReadGroups

BuildBamIndex

RealignerTargetCreator

IndelRealigner

MarkDuplicates

BuildBamIndex

BaseRecalibrator

PrintReads

bwa sampe

AddOrReplaceReadGroups

BuildBamIndex

RealignerTargetCreator

IndelRealigner

MarkDuplicates

BuildBamIndex

BaseRecalibrator

PrintReads

Note: Reference genome only needs to be indexed once

NA06984.realign.dedup.recal.bam

NA06984.realign.dedup.recal.bam

Genotyping

HaplotypeCaller

HaplotypeCaller

NA06984.g.vcf

Sample2.g.vcf

Sample3.g.vcf

Joint genotyping

GenotypeGVCFs

AllSamples.vcf

Filtering

VariantFiltration

AllSamples.filtered.vcf

View in IGV

Naming conventions

Initial file name according to information about the content

NA06984.ILLUMINA.low_coverage.17q

For each step of the exercise, create a new file

NA06984.ILLUMINA.low_coverage.17q.merge.bam

NA06984.ILLUMINA.low_coverage.17q.merge.realign.bam

NA06984.ILLUMINA.low_coverage.17q.merge.realign.dedup.bam

NA06984.ILLUMINA.low_coverage.17q.merge.realign.dedup.recal.bam

…

Regarding index files

Many steps in the exercise require that certain input files are indexed. For example the reference genome and the bam file.

Index files are usually NOT given as direct input to programs. The programs assume that index files are located in the same folder as the indexed input file.

Example:

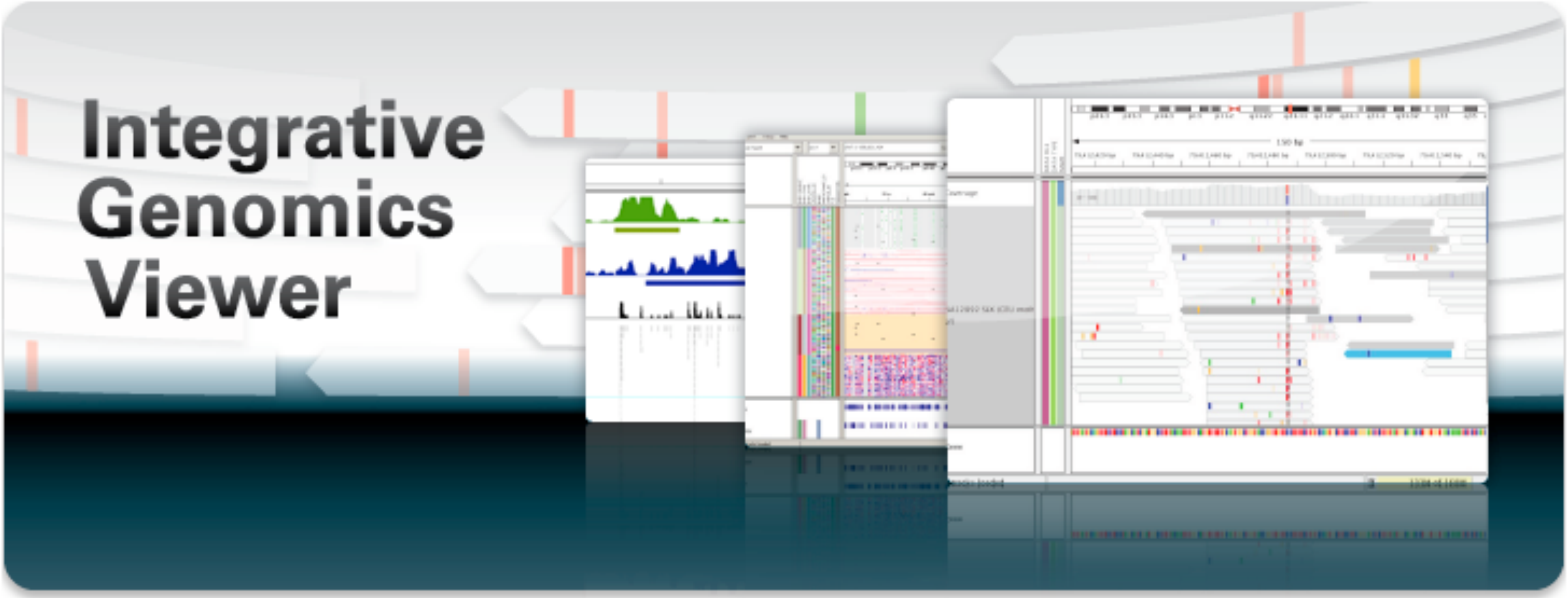
```
bwa sampe <ref> <sai1> <sai2> <fq1> <fq2> > align.sam
```

If you give the following file as reference:

~/glob/gatk/human_17_v37.fasta

BWA requires that index files exist in the folder ~/glob/gatk/

Viewing data with IGV

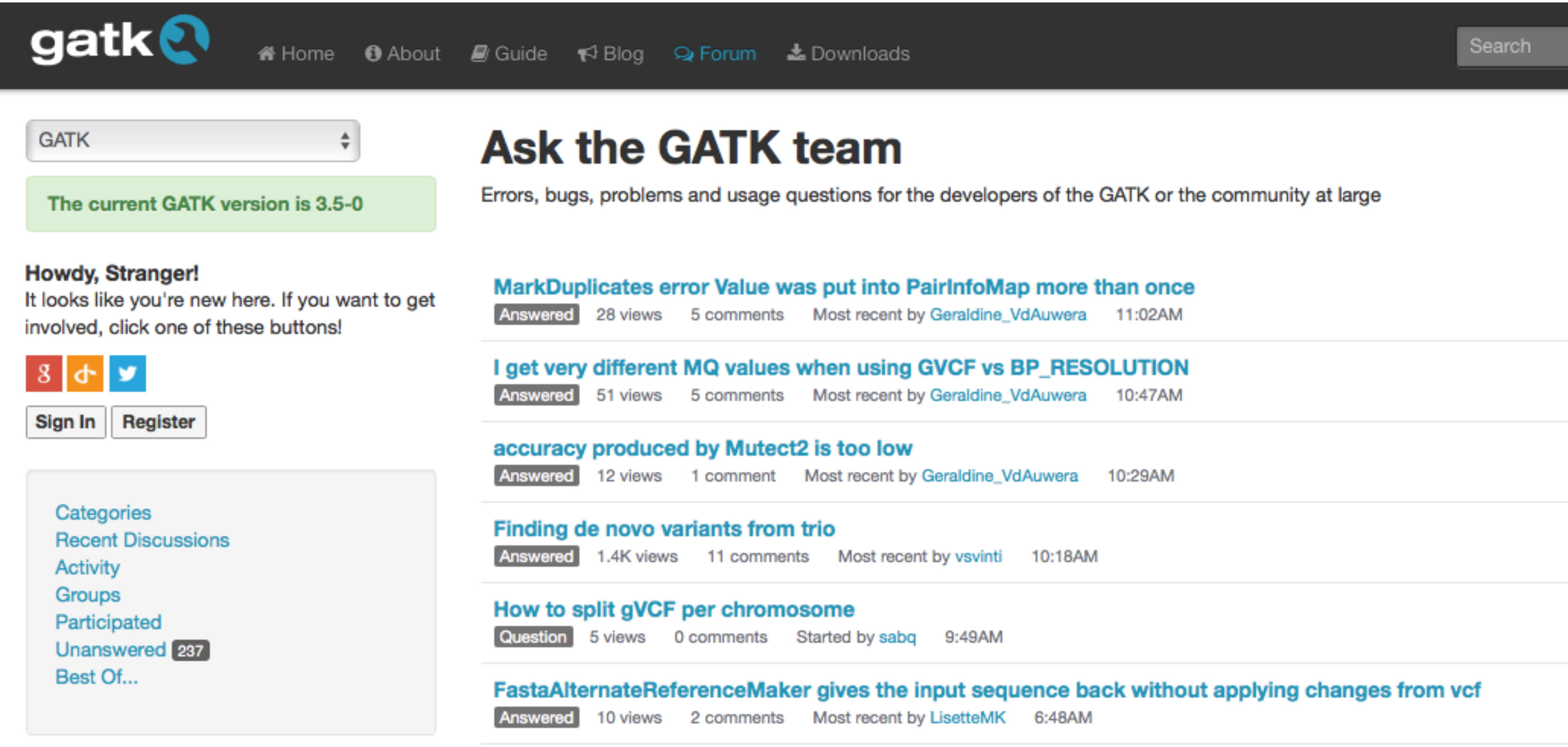


http://www.broadinstitute.org/igv/

GATK Support Forum

SciLifeLab

- https://www.broadinstitute.org/gatk/guide/best-practices
- https://www.broadinstitute.org/gatk/guide/tooldocs/
- http://gatkforums.broadinstitute.org/gatk/categories/ask-the-team



The current GATK version is 3.5-0

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

Ask the GATK team

Errors, bugs, problems and usage questions for the developers of the GATK or the community at large

MarkDuplicates error Value was put into PairInfoMap more than once
Answered 28 views 5 comments Most recent by Geraldine_VdAuwera 11:02AM

I get very different MQ values when using GVCF vs BP_RESOLUTION
Answered 51 views 5 comments Most recent by Geraldine_VdAuwera 10:47AM

accuracy produced by Mutect2 is too low
Answered 12 views 1 comment Most recent by Geraldine_VdAuwera 10:29AM

Finding de novo variants from trio
Answered 1.4K views 11 comments Most recent by vsvinti 10:18AM

How to split gVCF per chromosome
Question 5 views 0 comments Started by sabq 9:49AM

FastaAlternateReferenceMaker gives the input sequence back without applying changes from vcf
Answered 10 views 2 comments Most recent by LisetteMK 6:48AM

Index reference genome

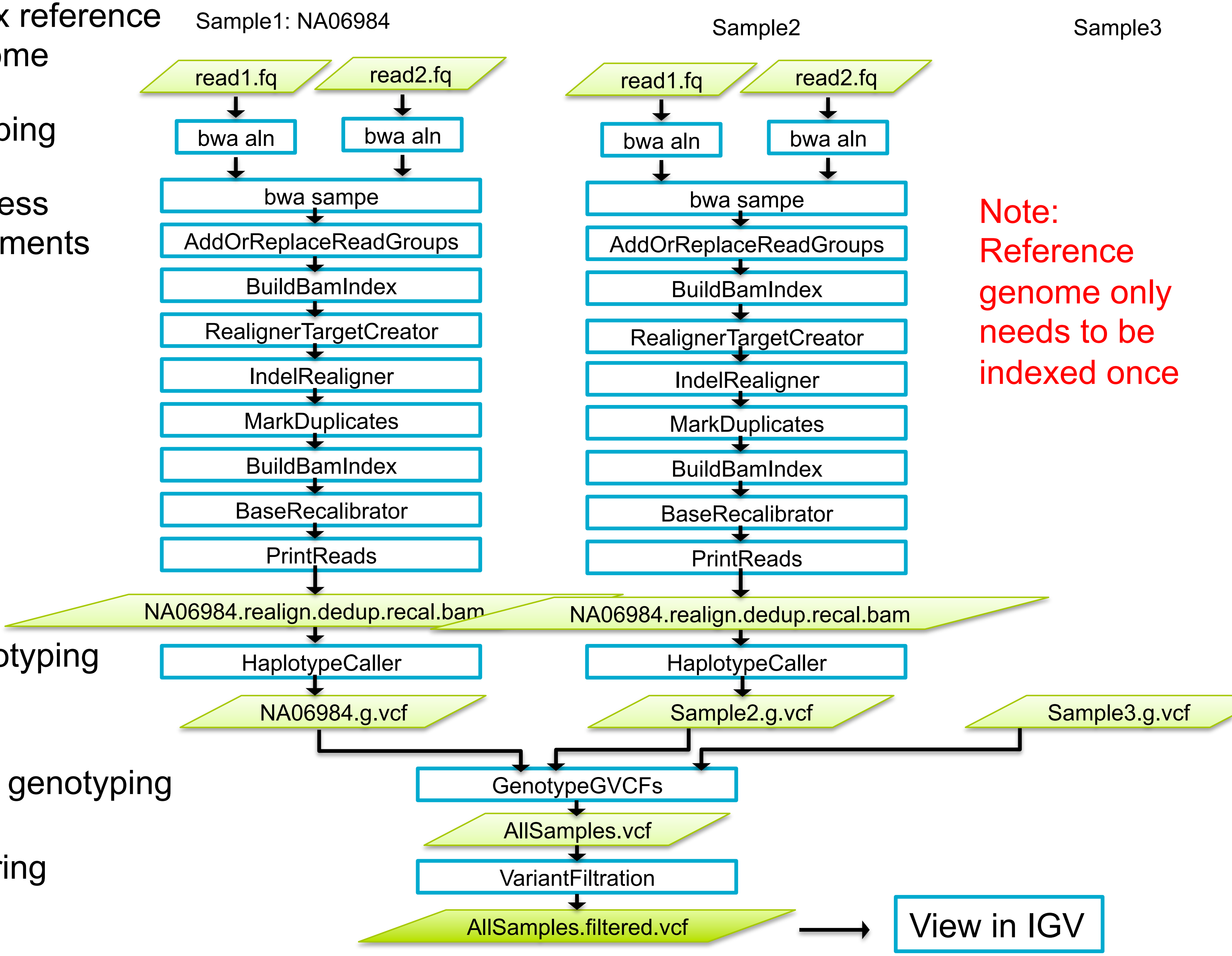
Mapping

Process alignments

Genotyping

Joint genotyping

Filtering



Sample1: NA06984

read1.fq

read2.fq

bwa aln

bwa aln

bwa sampe

AddOrReplaceReadGroups

BuildBamIndex

RealignerTargetCreator

IndelRealigner

MarkDuplicates

BuildBamIndex

BaseRecalibrator

PrintReads

NA06984.realign.dedup.recal.bam

HaplotypeCaller

NA06984.g.vcf

Sample2

read1.fq

read2.fq

bwa aln

bwa aln

bwa sampe

AddOrReplaceReadGroups

BuildBamIndex

RealignerTargetCreator

IndelRealigner

MarkDuplicates

BuildBamIndex

BaseRecalibrator

PrintReads

NA06984.realign.dedup.recal.bam

HaplotypeCaller

Sample2.g.vcf

Sample3

Sample3.g.vcf

GenotypeGVCFs

AllSamples.vcf

VariantFiltration

AllSamples.filtered.vcf

View in IGV

Note: Reference genome only needs to be indexed once

2) Align each paired end separately

SciLifeLab

```
bwa aln <ref> <fq1> > <sai1>
bwa aln <ref> <fq2> > <sai2>
```

<ref>	= reference sequence
<fq1>	= fastq reads seq 1 of pair
<fq2>	= fastq reads seq 2 of pair
<sai1>	= alignment of seq 1 of pair
<sai2>	= alignment of seq 2 of pair

3) Merging alignments

Combine alignments from paired ends into a SAM file

```
bwa sampe <ref> <sai1> <sai2> <fq1> <fq2> > align.sam
```

`<ref>`	= reference sequence
`<sai1>`	= alignment of seq 1 of pair
`<sai2>`	= alignment of seq 2 of pair
`<fq1>`	= fastq reads seq 1 of pair
`<fq2>`	= fastq reads seq 2 of pair

4) Creating and editing BAM files

- Create .bam and add read groups (picard)

```
java -Xmx2g –jar /path/AddOrReplaceReadGroups.jar
INPUT=<sam file>
OUTPUT=<bam file>
... more options
```

- index new BAM file (picard)

```
java -Xmx2g –jar /path/BuildBamIndex.jar
INPUT=<bam file>
... more options
```

5) Process BAM

- mark problematic indels (GATK)

```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-I <bam file>
-R <ref file>
-T RealignerTargetCreator
-o <intervals file>
```

- realign around indels (GATK)

```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-I <bam file>
-R <ref file>
-T IndelRealigner
-o <realigned bam>
-targetIntervals <intervals file>
```

5) Process BAM cont.

- mark duplicates (picard)

```
java -Xmx2g -jar /path/MarkDuplicates.jar
INPUT=<input bam>
OUTPUT=<marked bam>
METRICS_FILE=<metrics file>
```

- quality recalibration - compute covariation (GATK)

```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-T BaseRecalibrator
-I <input bam>
-R <ref file>
-knownSites <vcf file>
-recalFile <calibration table>
```

- Second step quality recalibration - compute covariation (GATK)

```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-T PrintReads -BQSR <calibration table>
-I <input bam>
-R <ref file>
-o <recalibrated bam>
```

6) Variant calling

- HaplotypeCaller (GATK)

```
java -Xmx2g
-jar /path/GenomeAnalysisTK.jar
-T HaplotypeCaller
-R <ref file>
-I <bam>
-o <filename.g.vcf>
-emitRefConfidence GVCF
-variant_index_type LINEAR
-variant_index_parameter 128000
```

Processing files

NEXT:

repeat steps 2-5 for at least another sample!

6) Genotyping gvcf

- Assigning genotypes based on joint analysis of multiple samples

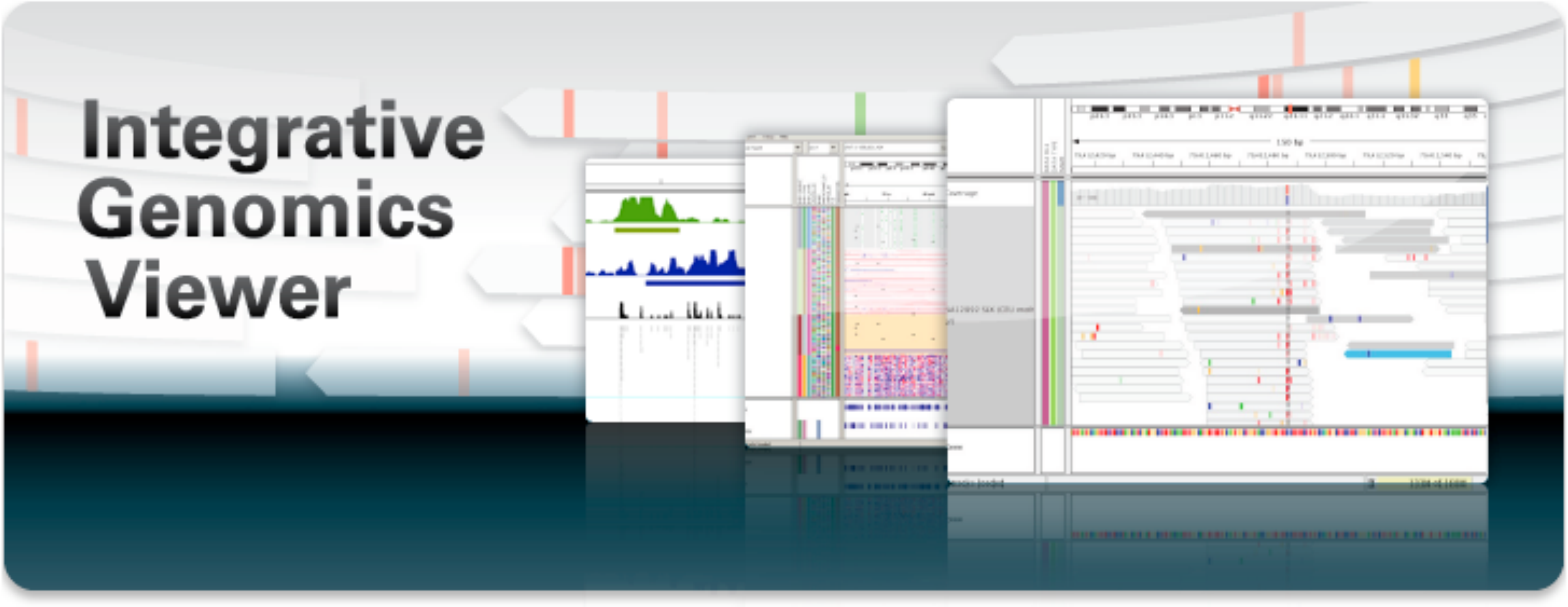
```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-T GenotypeGVCFs
-R <ref file>
--variant <sample1>.g.vcf
--variant <sample2>.g.vcf
...
-o <output vcf>
```

6) Filtering variants

- variant filtering

```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-T VariantFiltration
-R <reference>
-V <input vcf>
-o <output vcf>
--filterExpression "QD<2.0" --filterName QDfilter
--filterExpression "MQ<40.0" --filterName MQfilter
--filterExpression "FS>60.0" --filterName FSfilter
--filterExpression "HaplotypeScore>13.0" --filterName HSfilter
--filterExpression "MQRankSum<-12.5" --filterName MQRSfilter
--filterExpression "ReadPosRankSum<-8.0" --filterName RPRSfilter
```

7) Viewing data with IGV



http://www.broadinstitute.org/igv/

X) Extra

Extra 1: View data in UCSC-browser

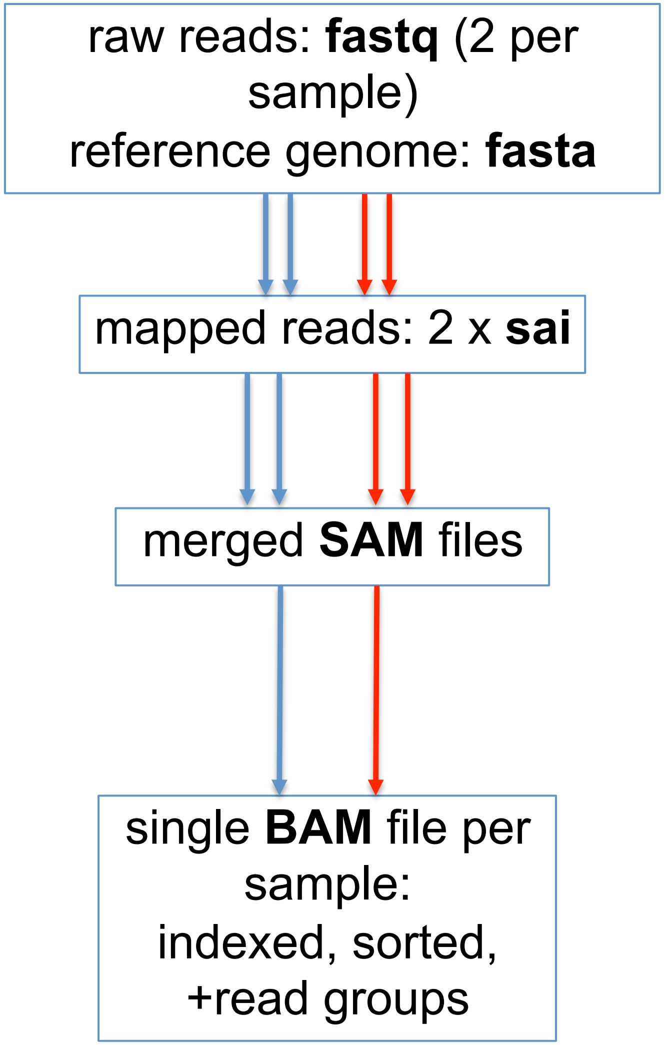
Extra 2: Select subset with BEDTools

Extra 3: Annotate variants with annovar

Extra 4: Make a script to run pipeline

pipeline (1)

2. Mapping
   - `bwa index`
   - `samtools faidx`
   - `bwa aln`
3. Merging alignments
   - `bwa sampe`
4. Creating BAM files
   - `picard AddOrReplaceReadGroups`
   - `picard BuildBamIndex`



raw reads: **fastq** (2 per sample)
reference genome: **fasta**

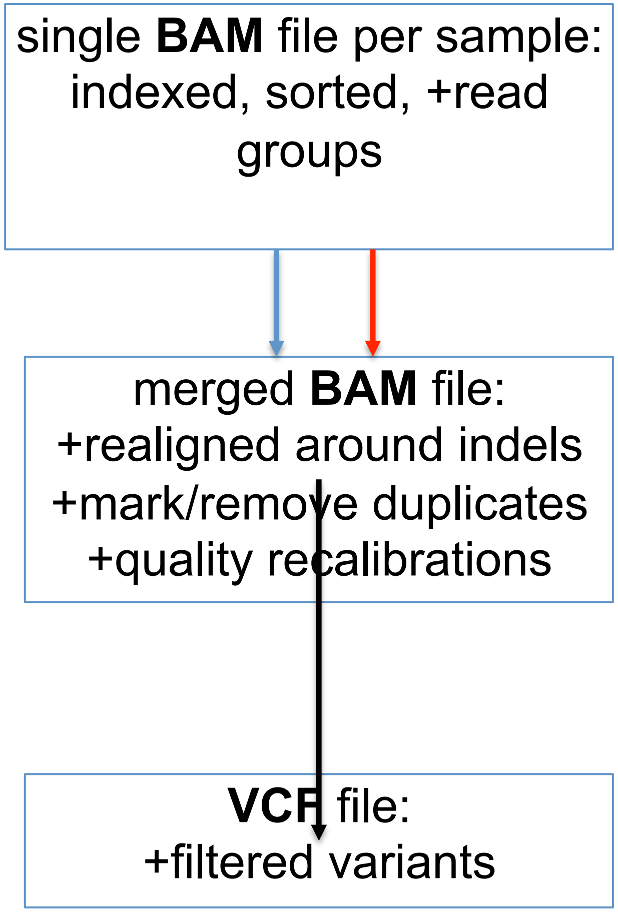
mapped reads: 2 x **sai**

merged **SAM** files

single **BAM** file per sample:
indexed, sorted,
+read groups

pipeline (2)

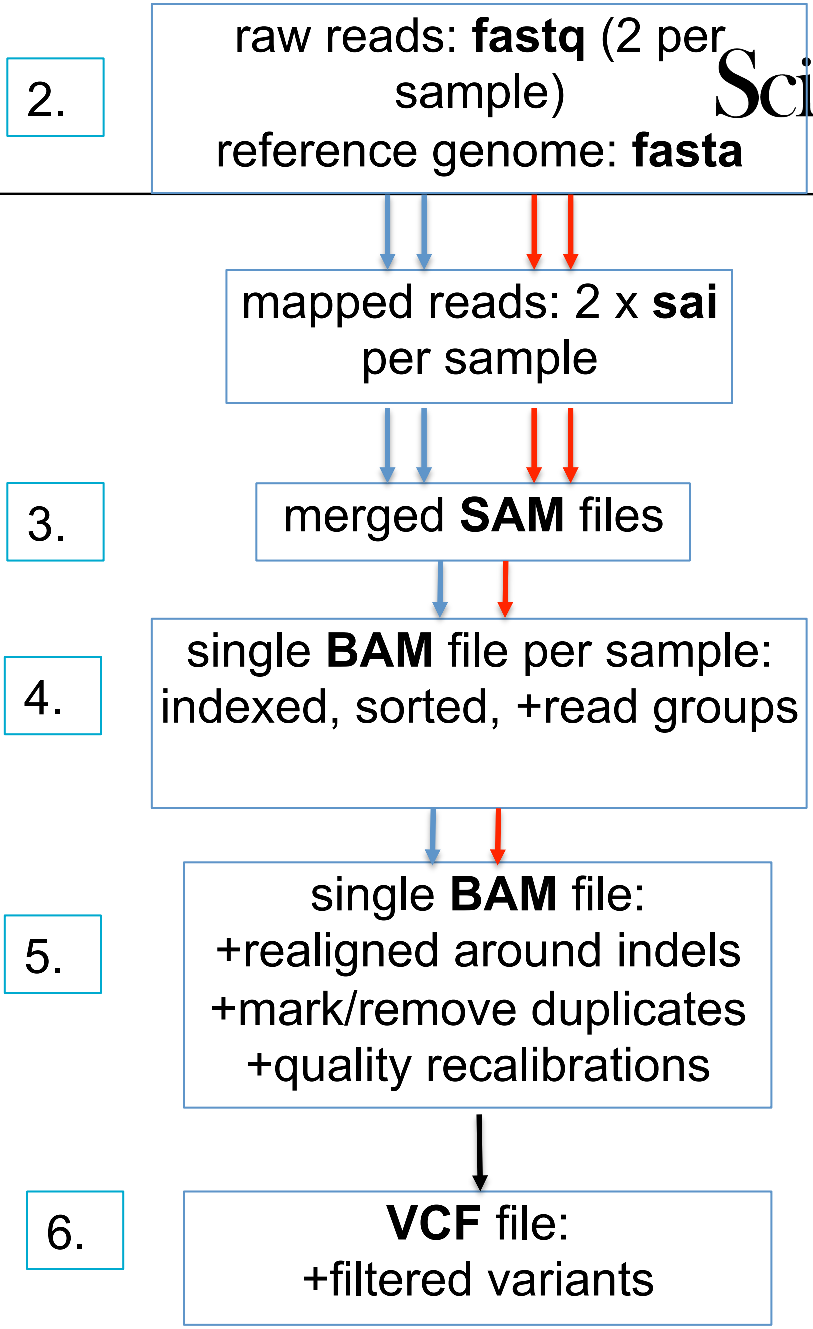
5. Processing files (GATK)
   - `GATK RealignerTargetCreator`
   - `GATK IndelRealigner`
   - `picard MarkDuplicates`
   - `GATK CountCovariates`
   - `picard MergeSamFiles`
6. Variant calling and filtering (GATK)
   - `GATK UnifiedGenotyper`
   - `GATK VariantFiltration`
7. Viewing data (IGV)



single **BAM** file per sample: indexed, sorted, +read groups

merged **BAM** file: +realigned around indels +mark/remove duplicates +quality recalibrations

VCF file: +filtered variants



2.

raw reads: **fastq** (2 per sample)
reference genome: **fasta**

mapping

mapped reads: 2 x **sai** per sample

3.

merged **SAM** files

4.

single **BAM** file per sample:
indexed, sorted, +read groups

processing

5.

single **BAM** file:
+realigned around indels
+mark/remove duplicates
+quality recalibrations

variant calling

6.

VCF file:
+filtered variants