SciLifeLab

Analysing re-sequencing samples

Malin Larsson

Malin.larsson@scilifelab.se

WABI / SciLifeLab

Re-sequencing

SciLifeLab

Reference genome assembly

...GTGCGTAGACTGCTAGATCGAAGA...

Re-sequencing

SciLifeLab

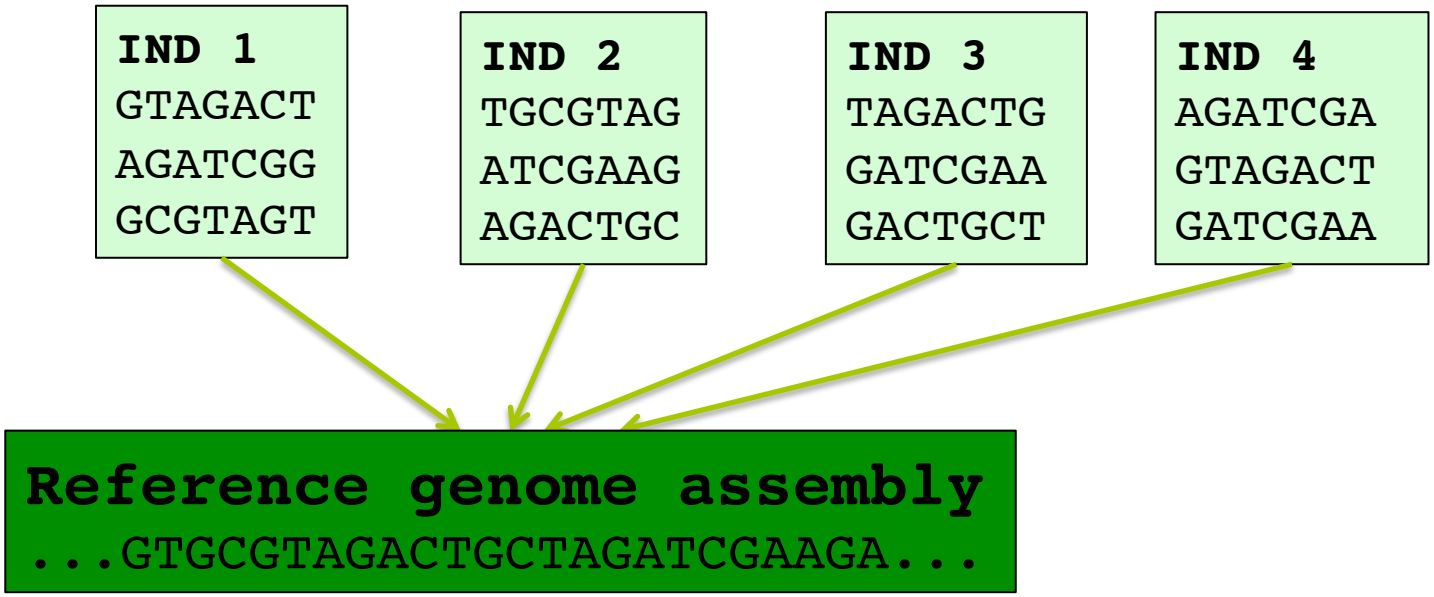
IND 1	IND 2	IND 3	IND 4
GTAGACT	TGCGTAG	TAGACTG	AGATCGA
AGATCGG	ATCGAAG	GATCGAA	GTAGACT
GCGTAGT	AGACTGC	GACTGCT	GATCGAA

Reference genome assembly

...GTGCGTAGACTGCTAGATCGAAGA...

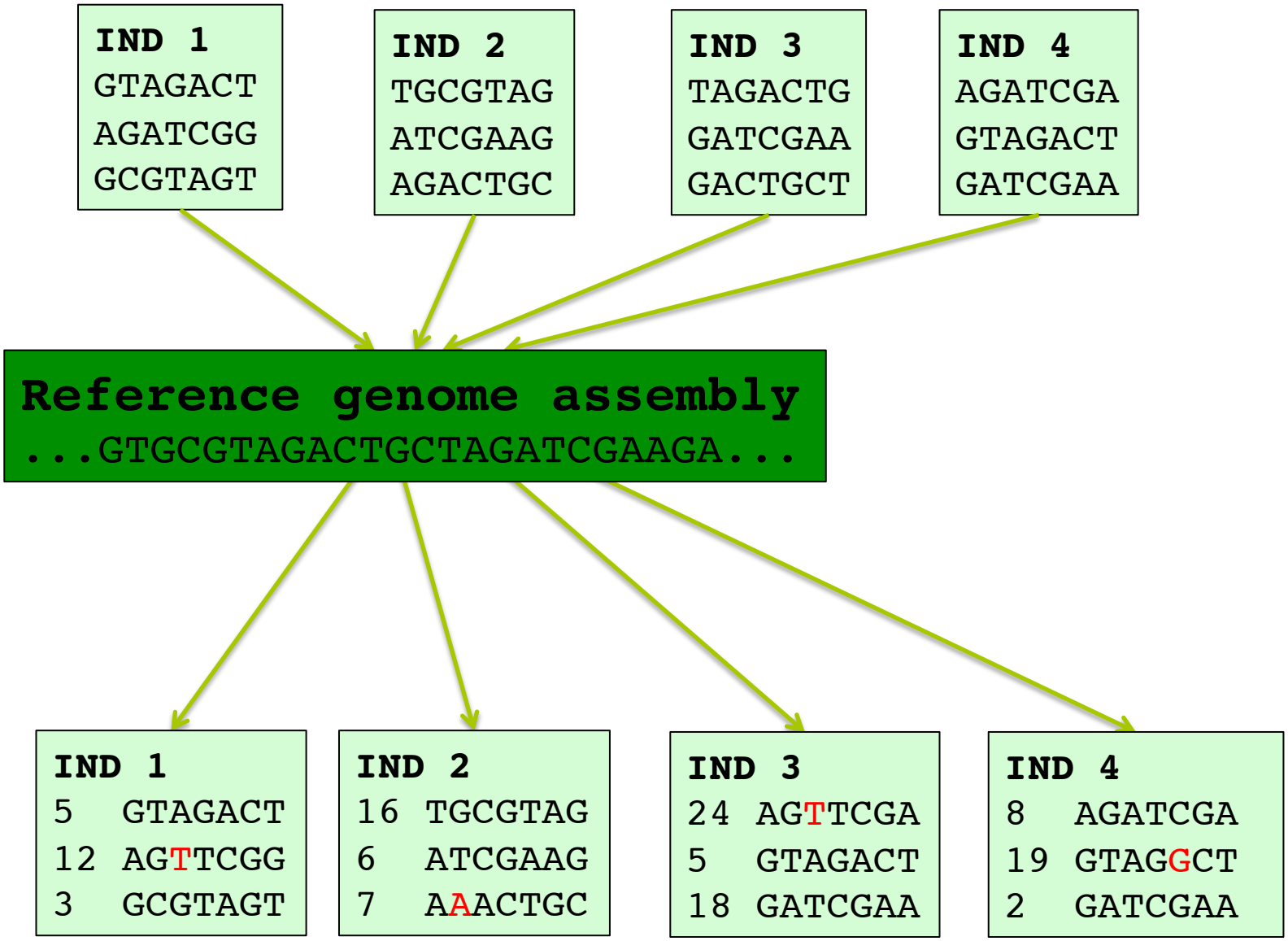
Re-sequencing

SciLifeLab



Re-sequencing

SciLifeLab



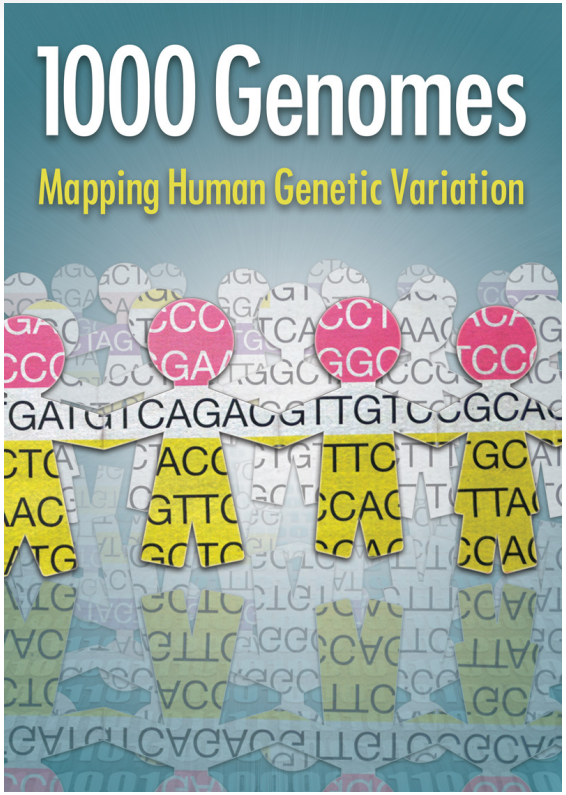
Rare variants in human

SciLifeLab

deCODE genetics

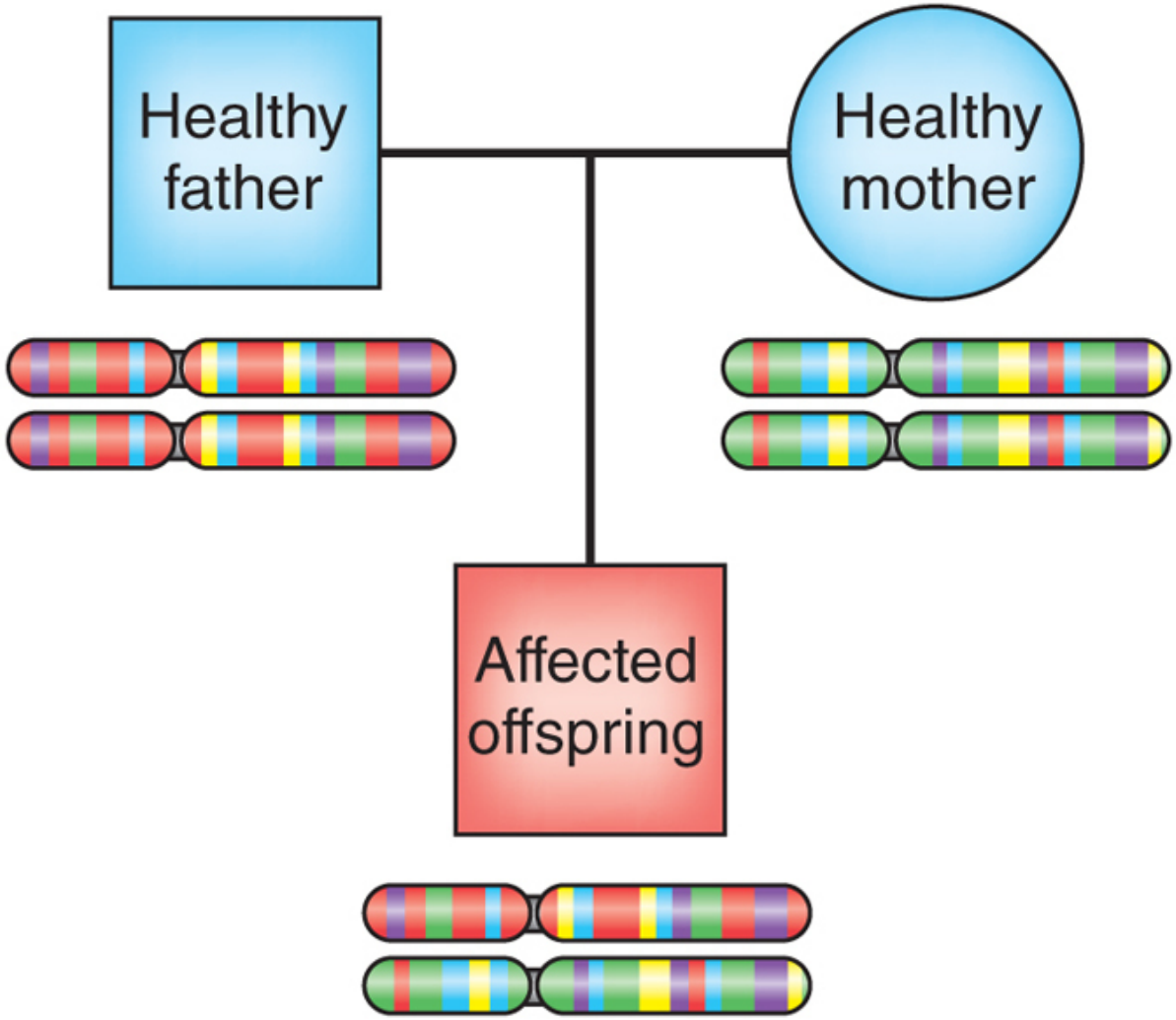
UK 10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE



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

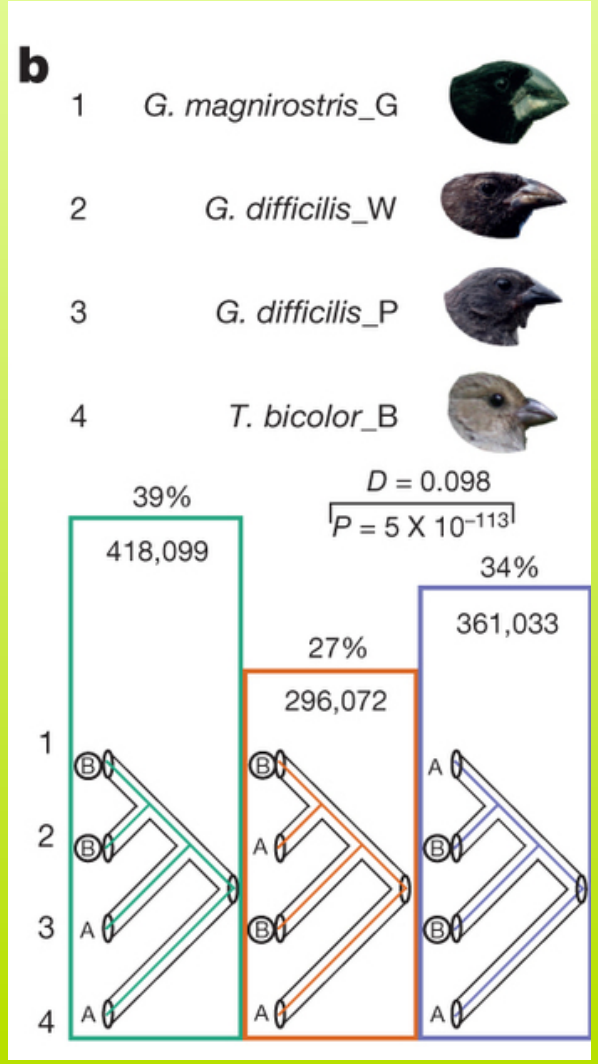
SciLifeLab



Population genetics – speciation, adaptive evolution

SciLifeLab

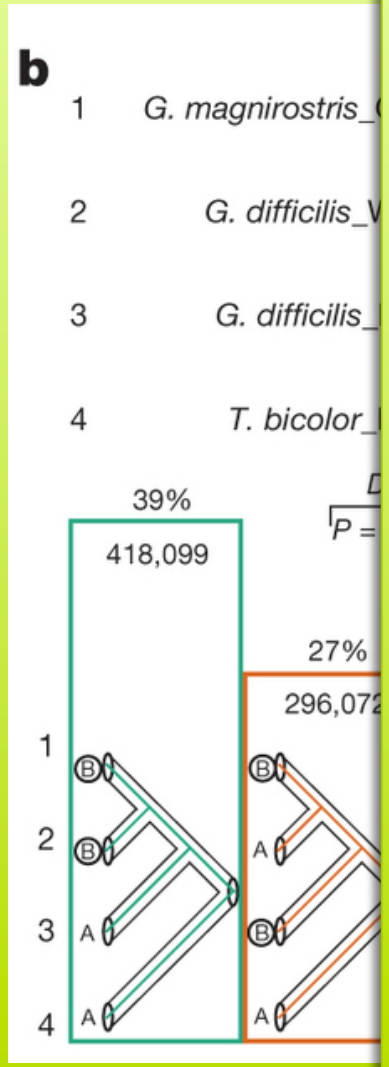
Darwin Finches



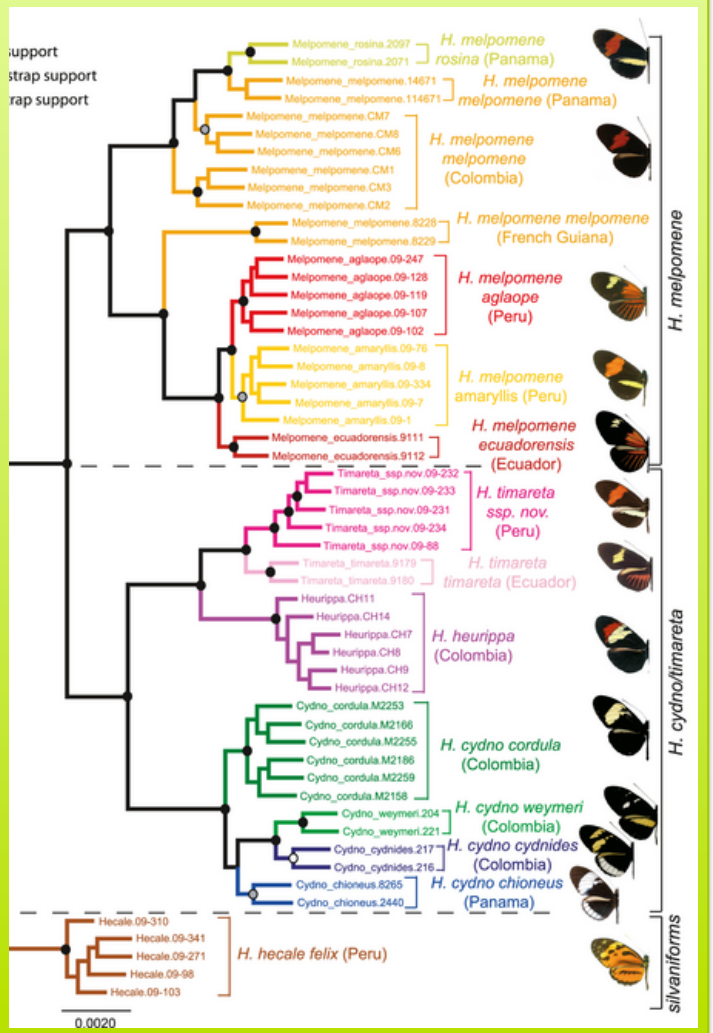
Population genetics – speciation, adaptive evolution

SciLifeLab

Darwin Finches



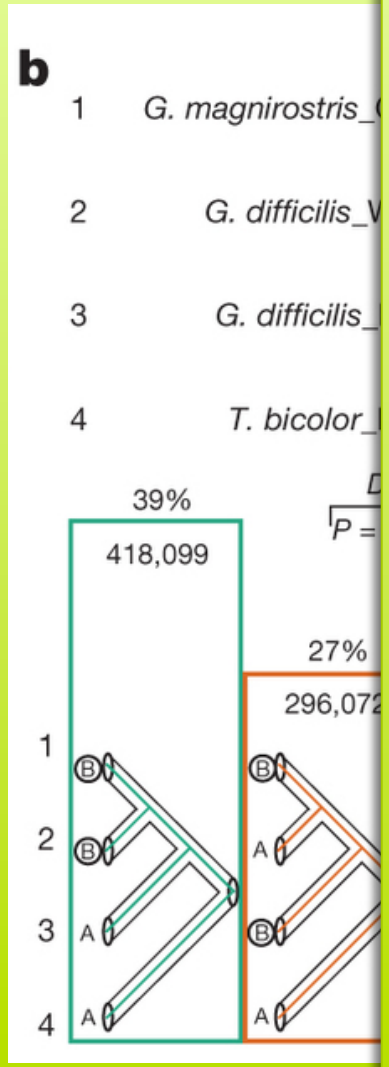
Heliconius Butterflies



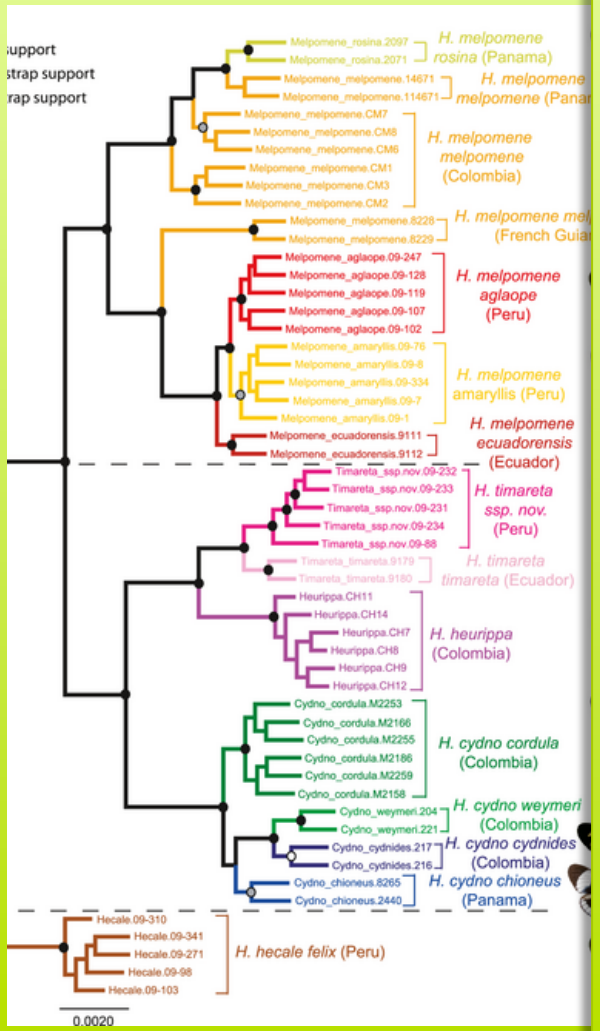
Population genetics – speciation, adaptive evolution

SciLifeLab

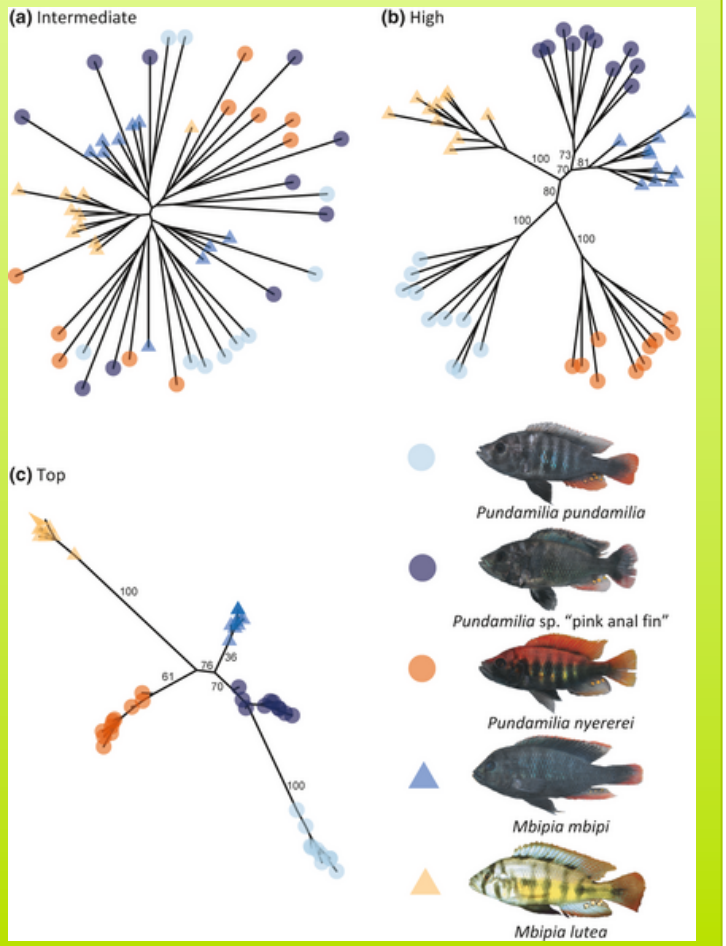
Darwin Finches



Heliconius Butterflies

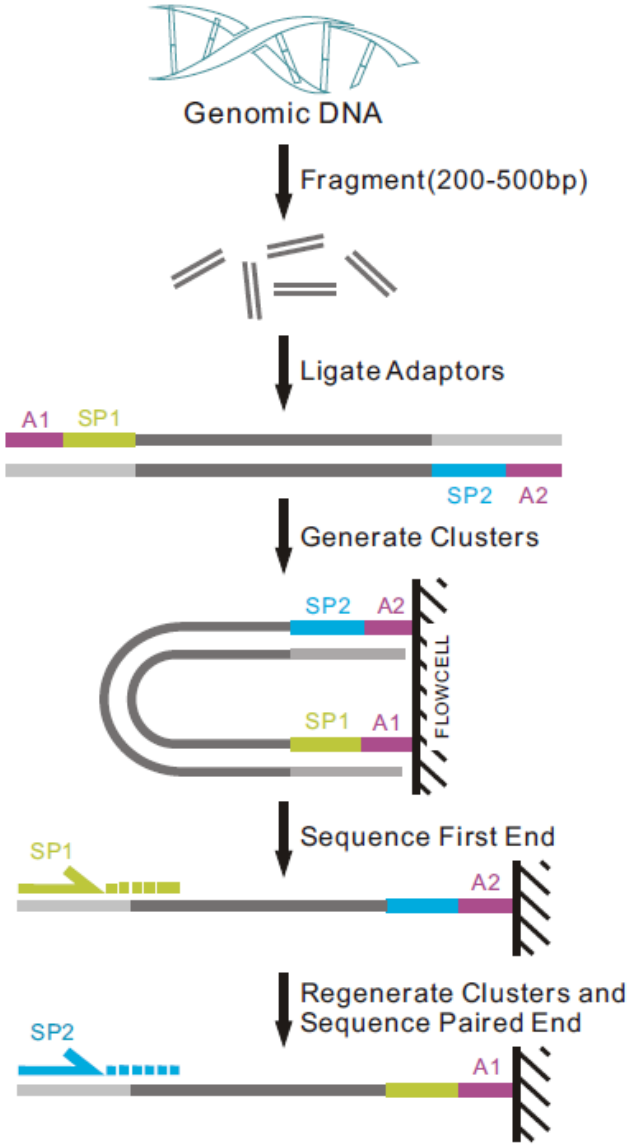


Lake Victoria cechlid fishes



Paired end sequencing

SciLifeLab



Pair-end reads

SciLifeLab

- 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@CFFFFFHHHHHGJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

ID_R2_001.fastq

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

Pair-end reads

SciLifeLab

- 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

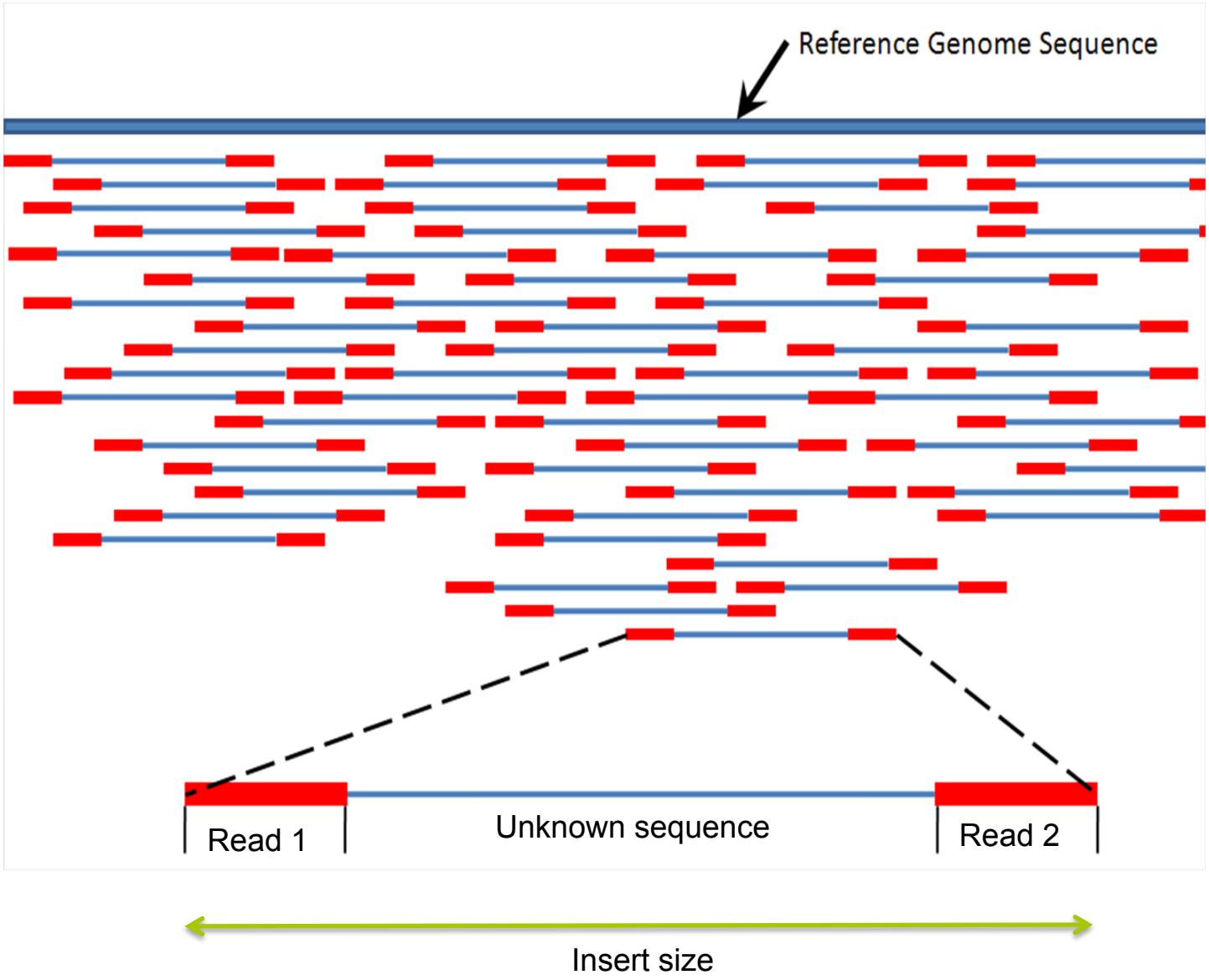
ID_R1_001.fastq

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

ID_R2_001.fastq

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

Paired end sequencing

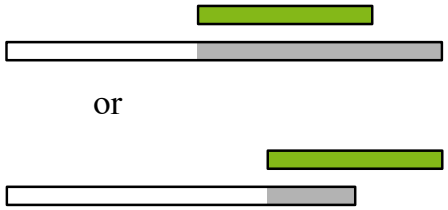


Adapter trimming

SciLifeLab

```
Module load cutadapt
```

3' Adapter



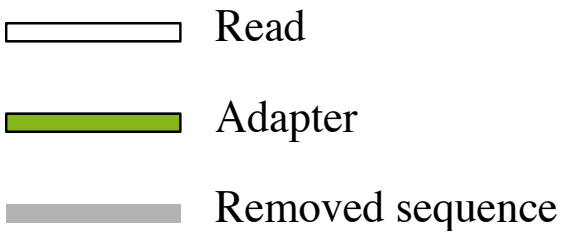
or

5' Adapter

or

Anchored 5' adapter

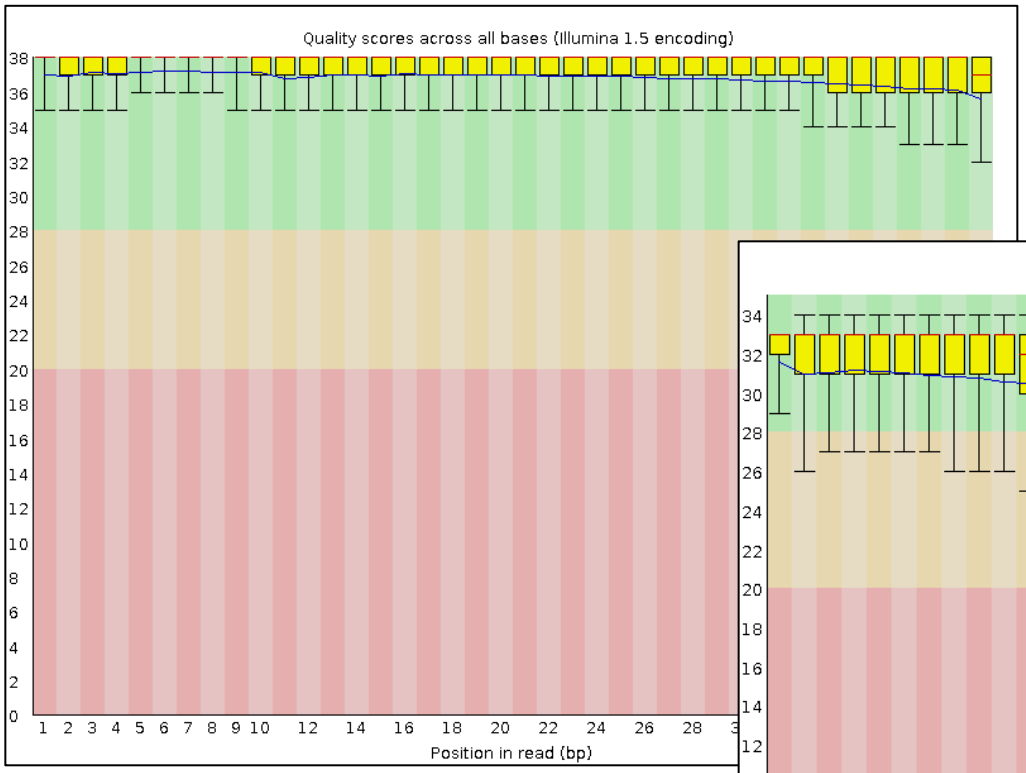
When the adaptor has been read in sequencing, it is present in reads and needs to be removed prior to mapping

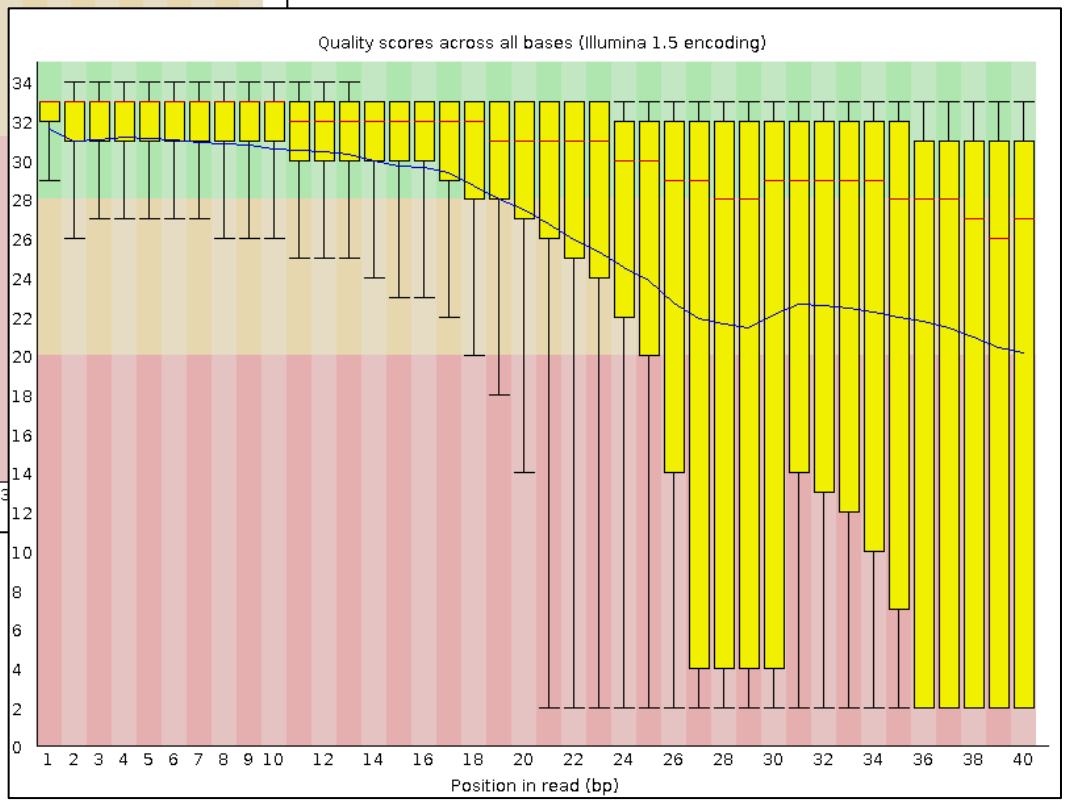


Basic quality control - FASTQC

SciLifeLab

```
Module load FastQC
```





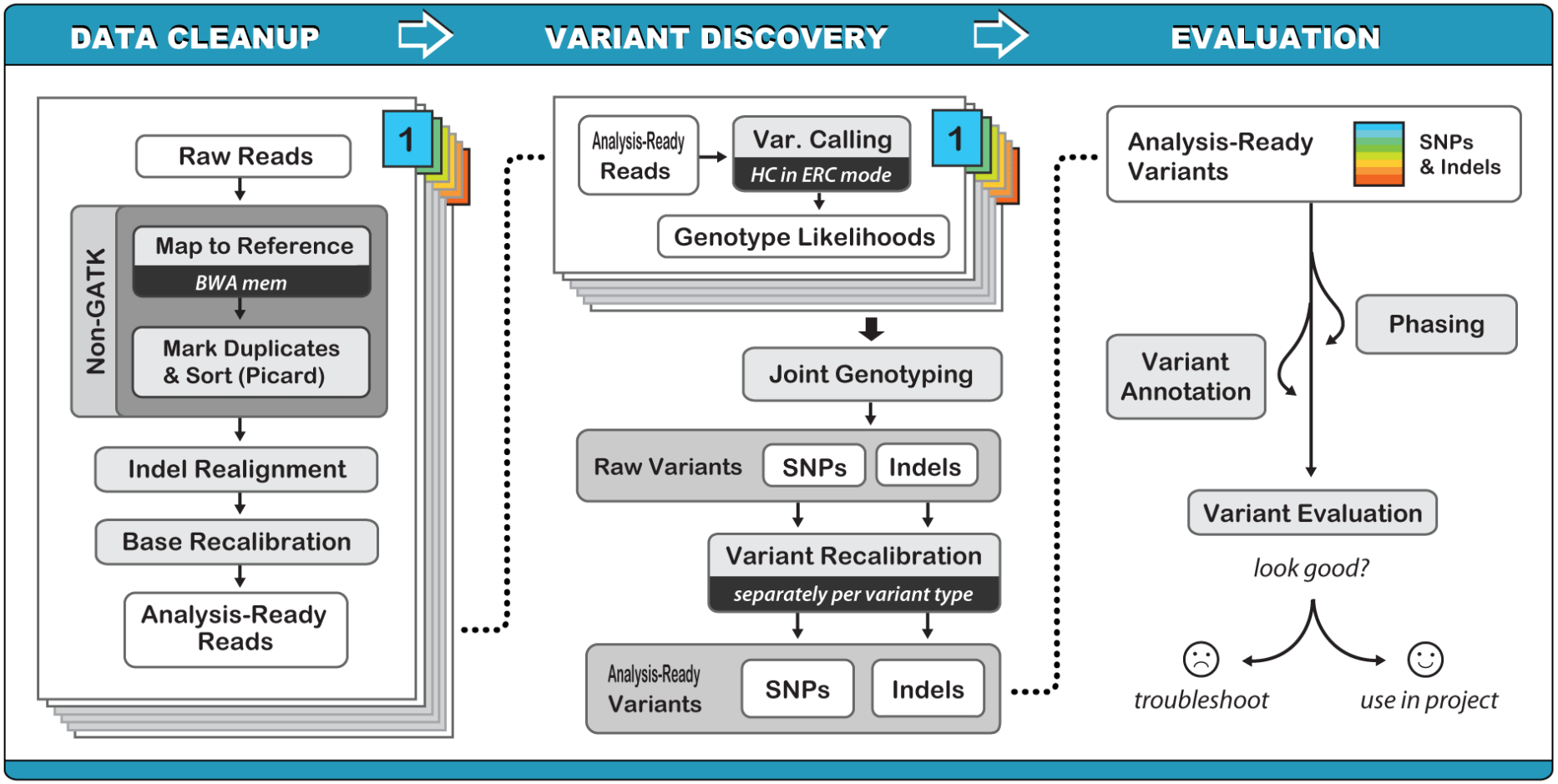
Genome Analysis Tool Kit (GATK)

SciLifeLab

Mapping Alignment refinement Variant discovery Callset refinement

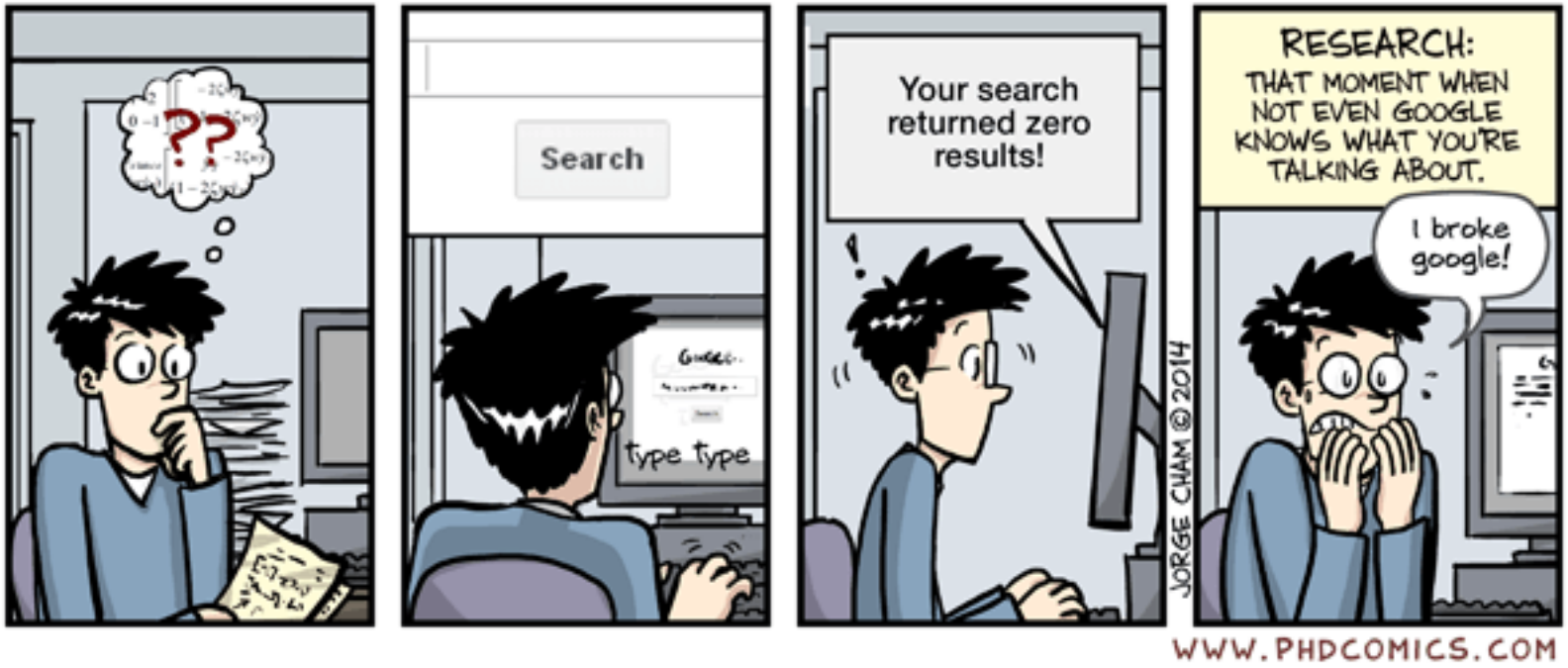
GATK

SciLifeLab



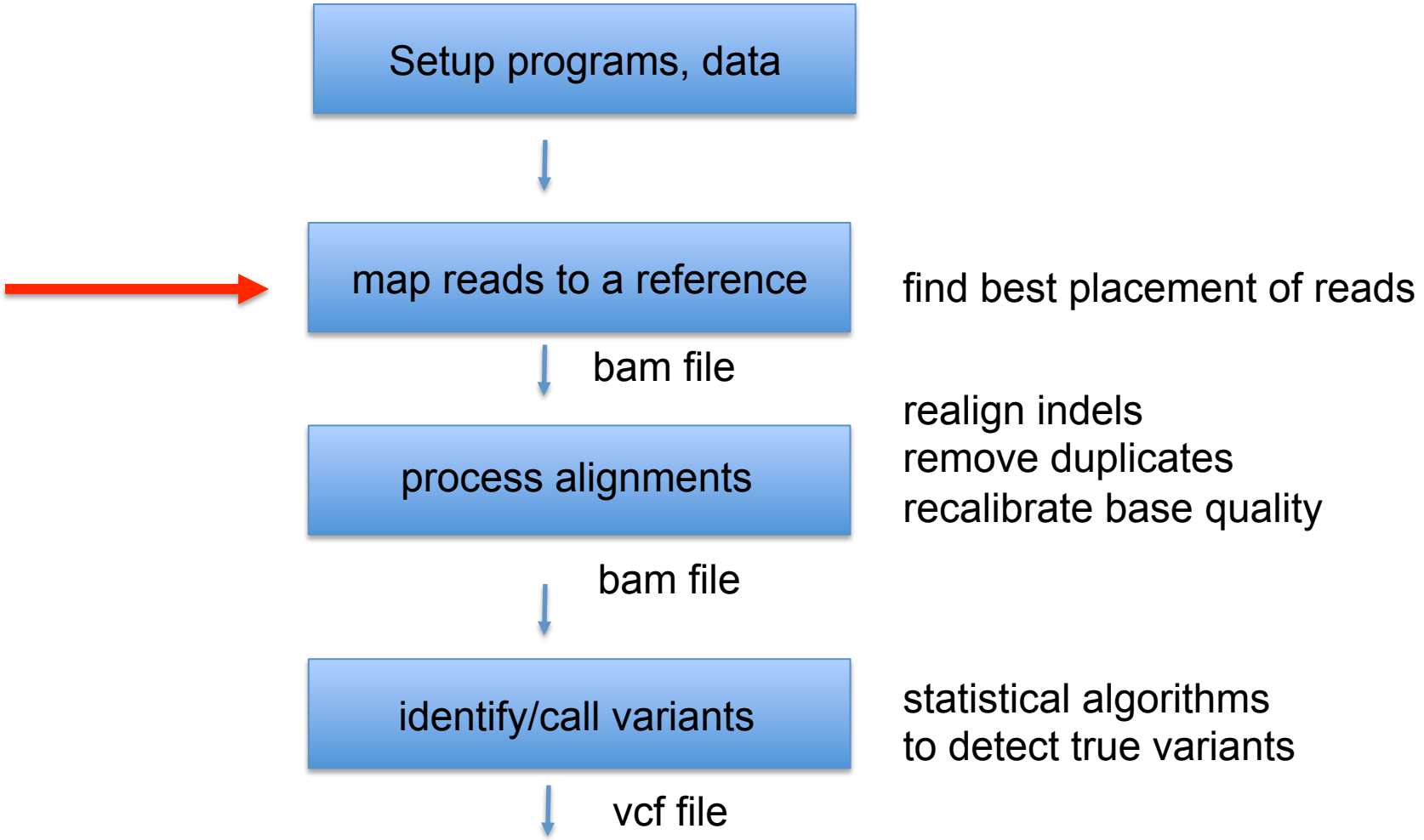
When in doubt, google it!

SciLifeLab



Steps in resequencing analysis

SciLifeLab



Mapping to reference genome

brute force

SciLifeLab

```
TCGATCC
x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
 TCGATCC
 x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
   TCGATCC
   x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
    TCGATCC
    x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
      TCGATCC
      ||x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
      TCGATCC
      x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

```
      TCGATCC
      x
GACCTCATCGATCCCACTG
```

brute force

SciLifeLab

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

hash tables

SciLifeLab

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

SciLifeLab

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

SciLifeLab

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

SciLifeLab

Transformation				
Input	All Rotations	Sorting All Rows in Alphabetical Order by their first letters	Taking Last Column	Output Last Column
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	**A**NANA|^B **AN**A|^BAN **A**|^BANAN **B**ANANA|^ **N**ANA|^BA **NA**|^BANA ^BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|^ NANA|^B**A** NA|^BAN**A** ^BANANA| |^BANAN**A**	BNN^AA|A

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

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

Input to mapping

SciLifeLab

Reference genome

Reference.fasta

Reference.fai

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

Sample data

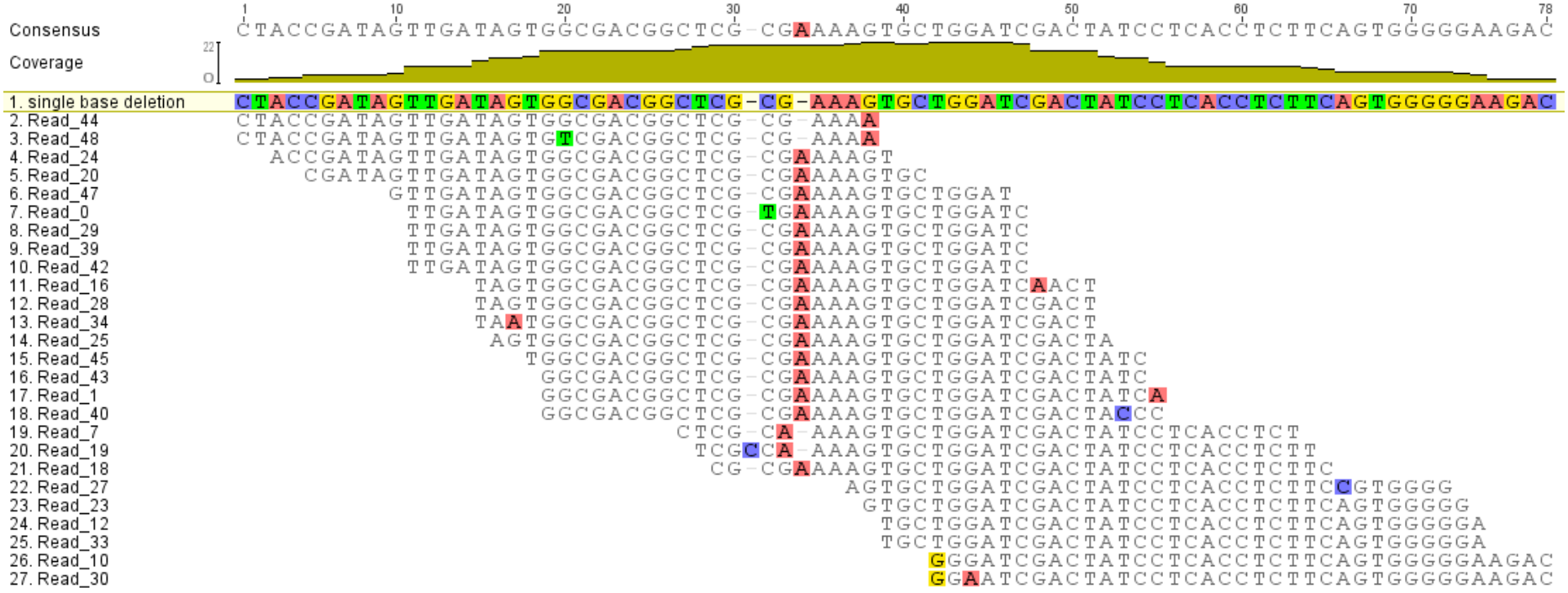
R1.fastq

R2.fastq

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

Output from mapping

SciLifeLab



Output - SAM format

SciLifeLab

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
AACATAAACACATCTCTATCAGAATTAGAGTAAAAGACCCCTGCCCCCAAGCAAAGGAT      >BA@>=@?<@@AA@A?@!@@;@AAB;A?AA@A<A<A<@?>A@@A@>?!=>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

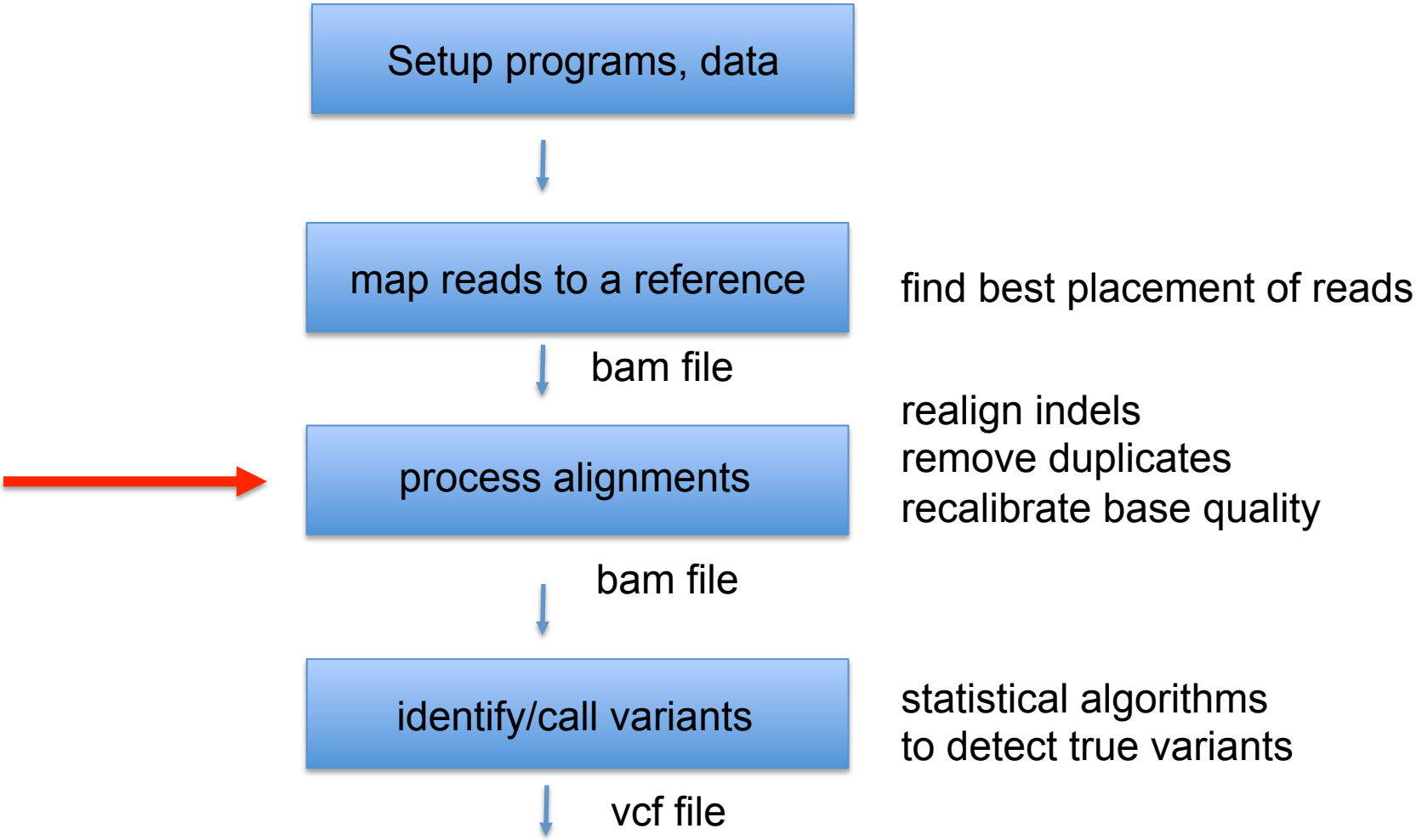
Start position

Quality

Sequence

Steps in resequencing analysis

SciLifeLab



Processing BAM files

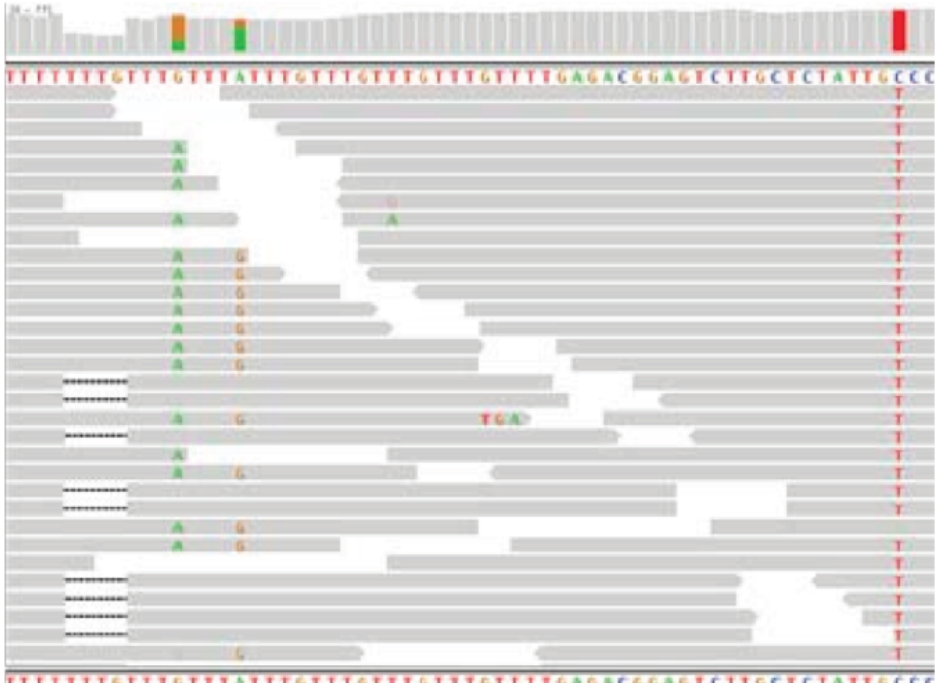
SciLifeLab

.bam

Realign around indels

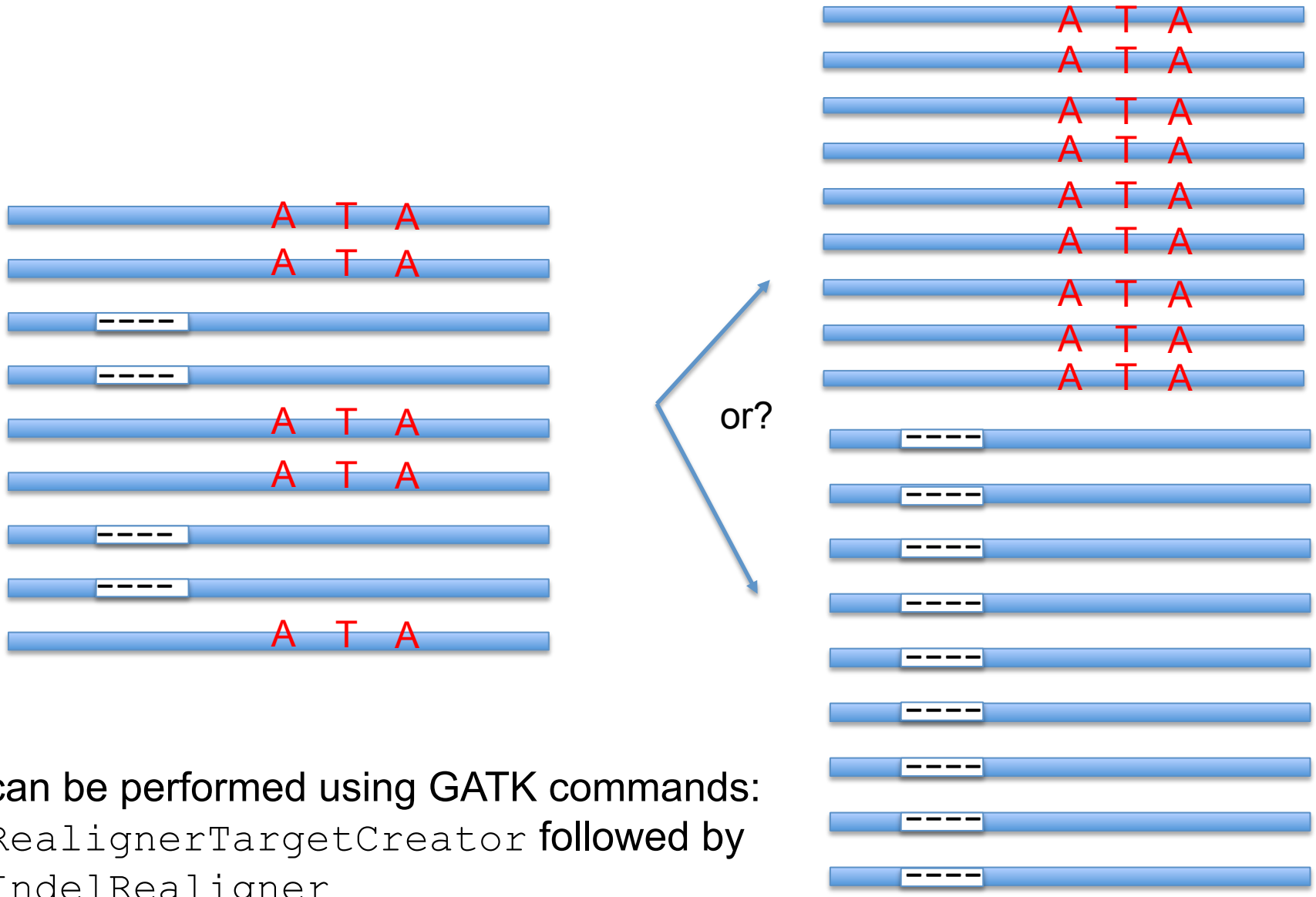
SciLifeLab

- 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



Local realignment

SciLifeLab



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

PCR duplicates

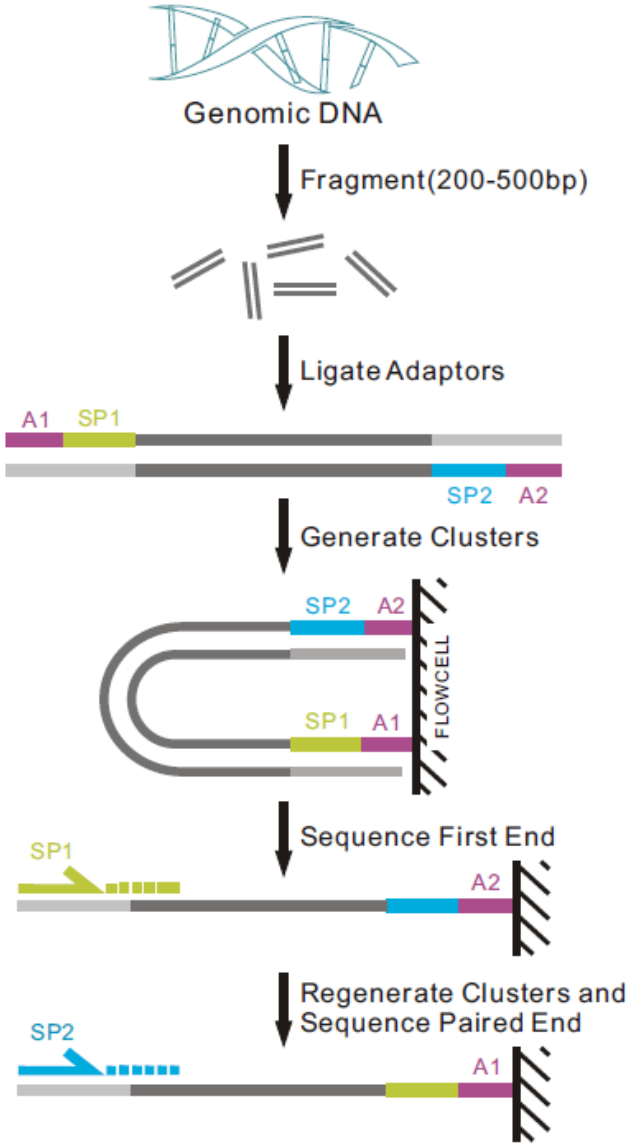
SciLifeLab

- When two or more reads originate from same molecule (artificial duplicates)
  - 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)
- mark or remove

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

Paired end sequencing

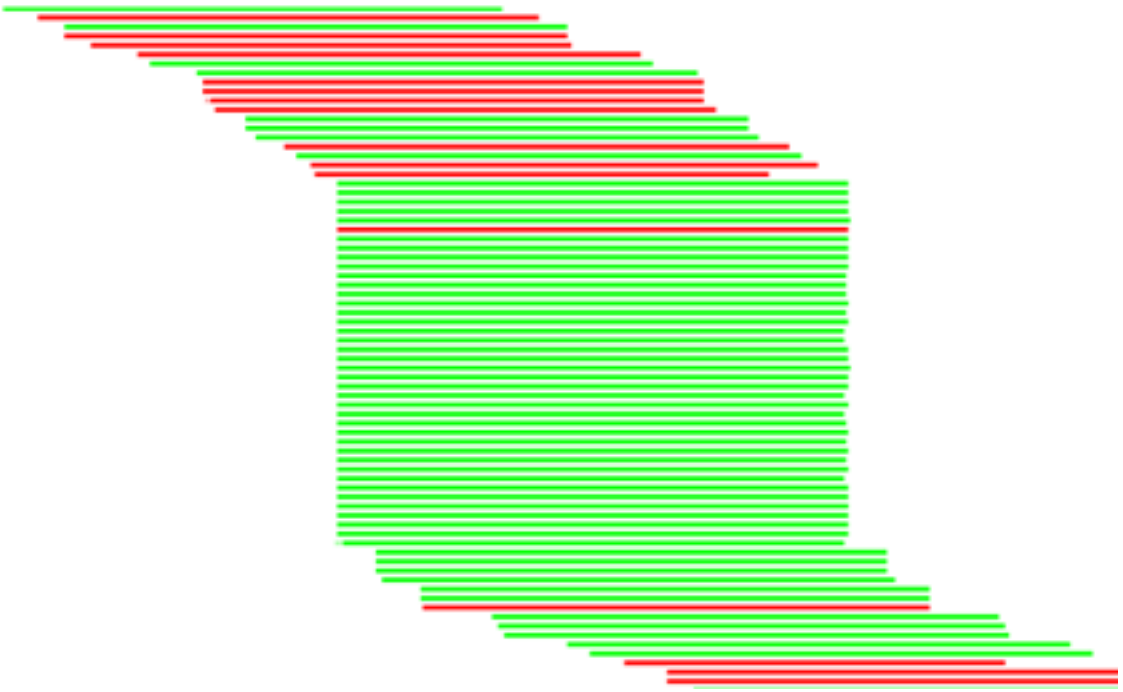
SciLifeLab



Identify PCR duplicates

SciLifeLab

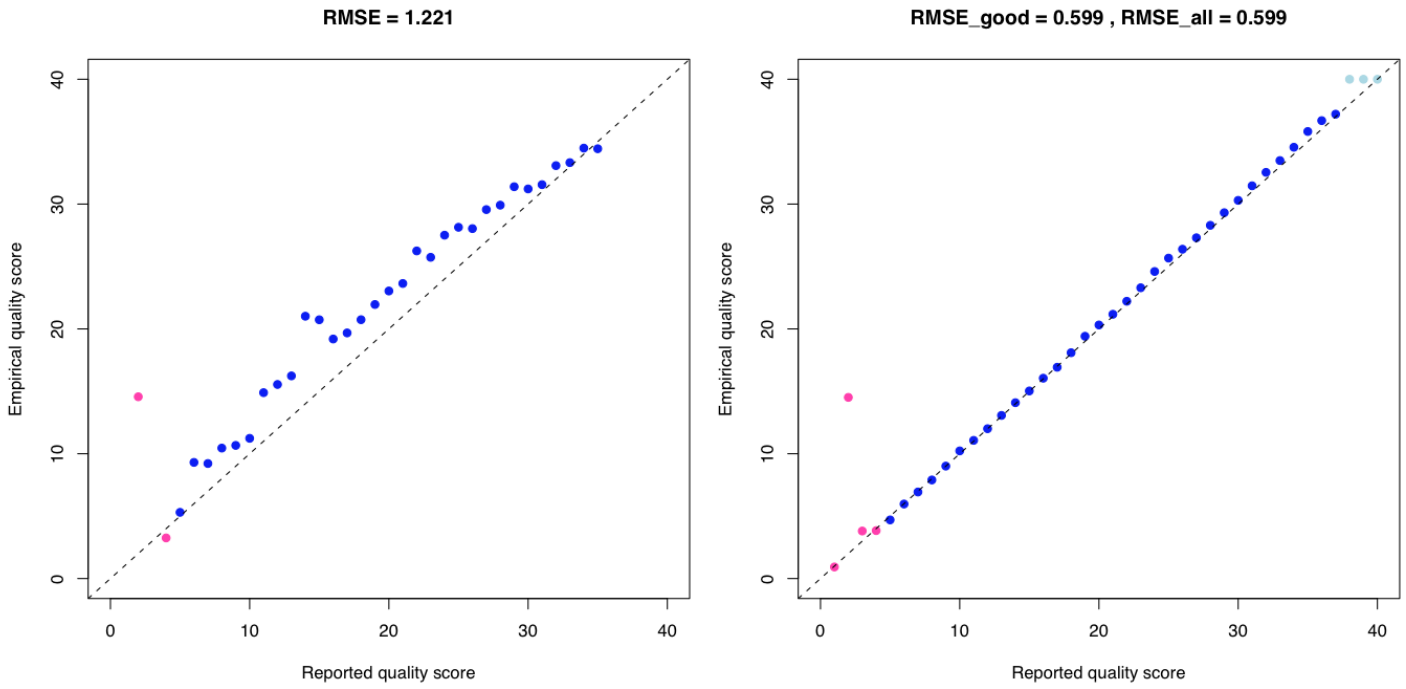
- Single or paired reads that map to identical positions
- Picard `MarkDuplicates`



Base quality recalibration

SciLifeLab

Reported Quality vs. Empirical Quality



Original Data

After GATK Recalibration

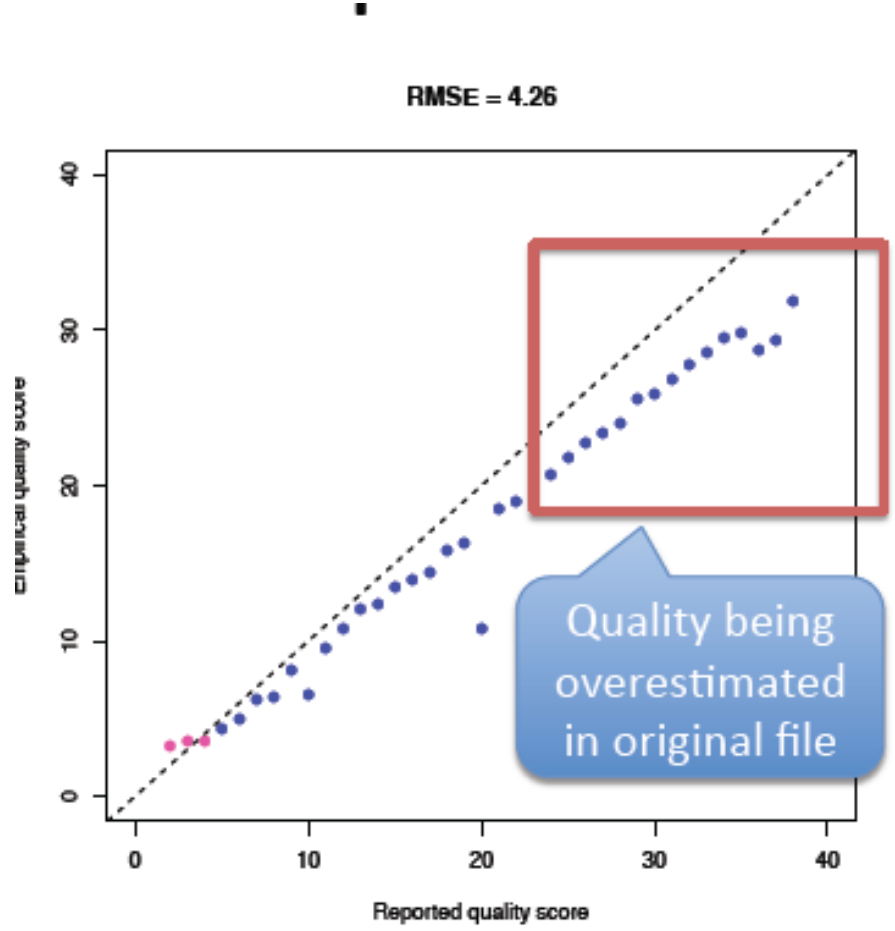
Recalibration Method

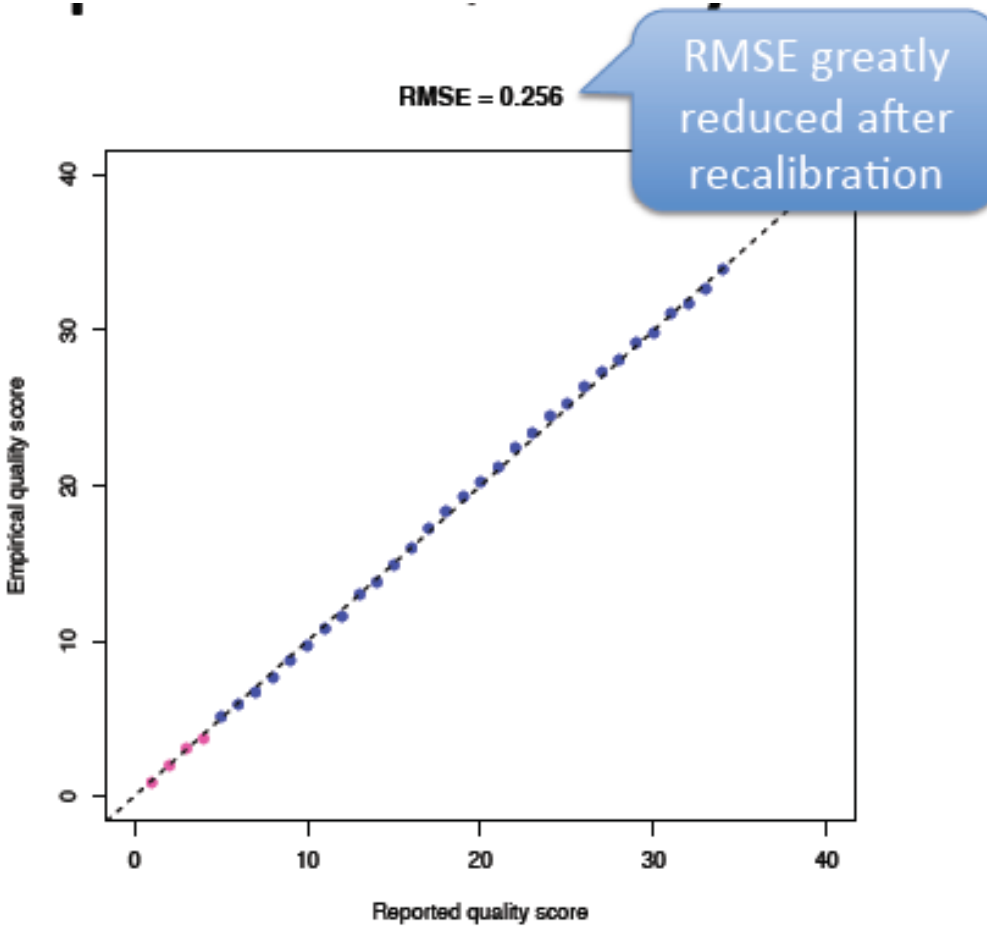
SciLifeLab

- Bin each base by
  - read group
  - called quality
  - position in read
  - local dinucleotide context
- score observed quality per bin
  - # of mismatches +1 / # of observed bases
- scale compared to reported quality
- Further reading:
- http://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr

Reported vs empiral quality scores

SciLifeLab



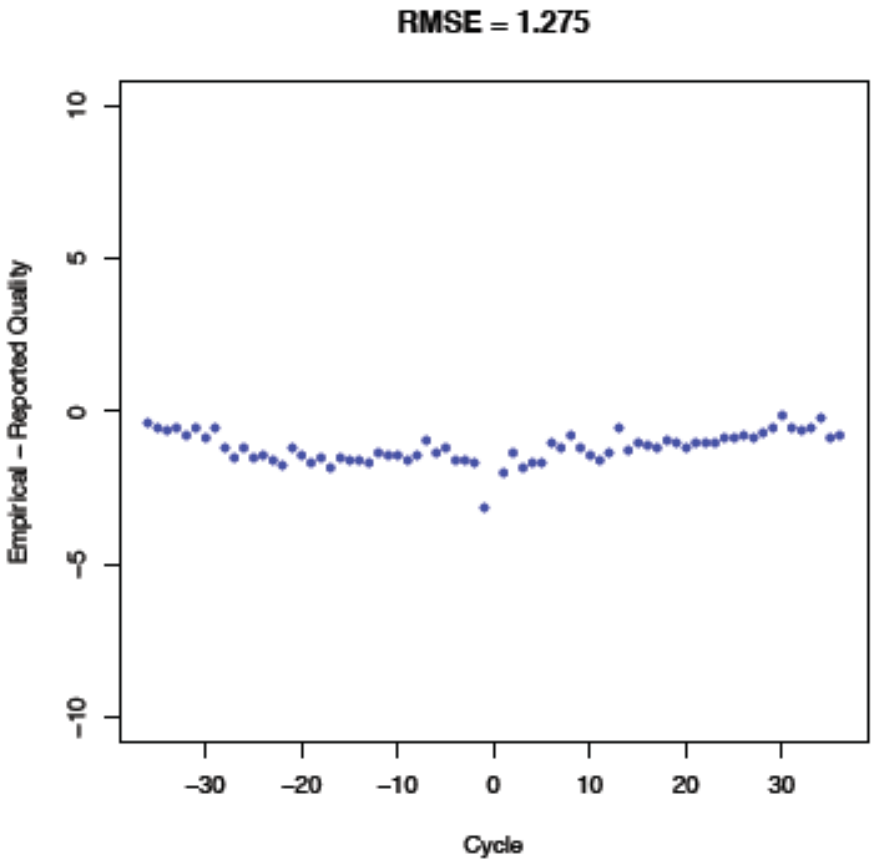


Before Recalibration

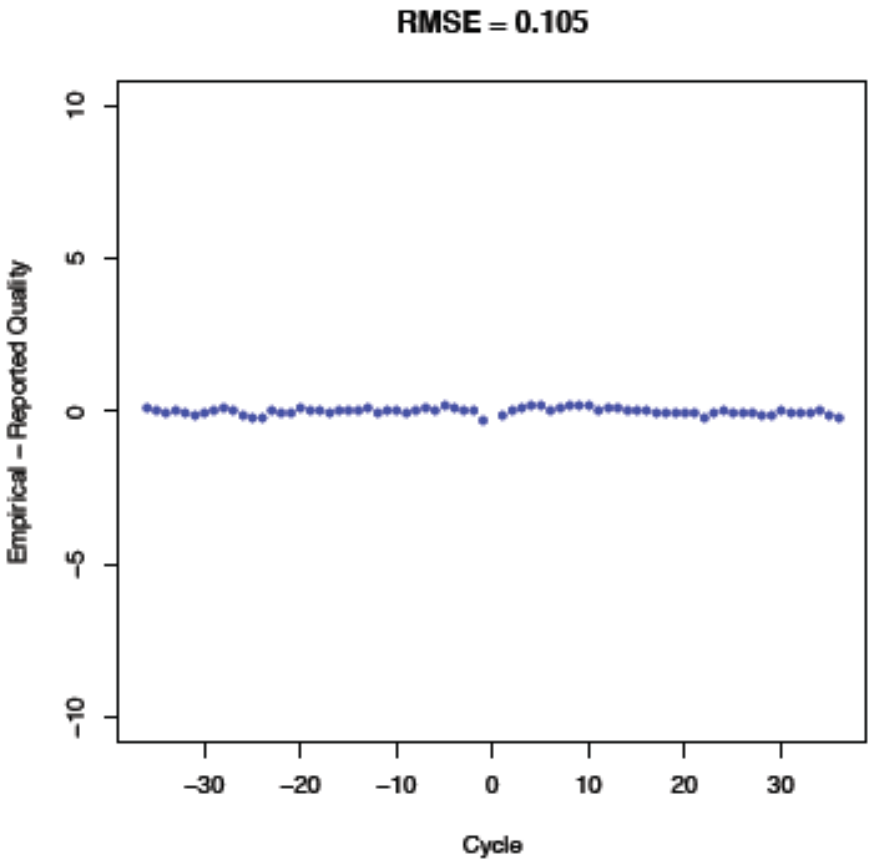
After Recalibration

Residual error by machine cycle

SciLifeLab



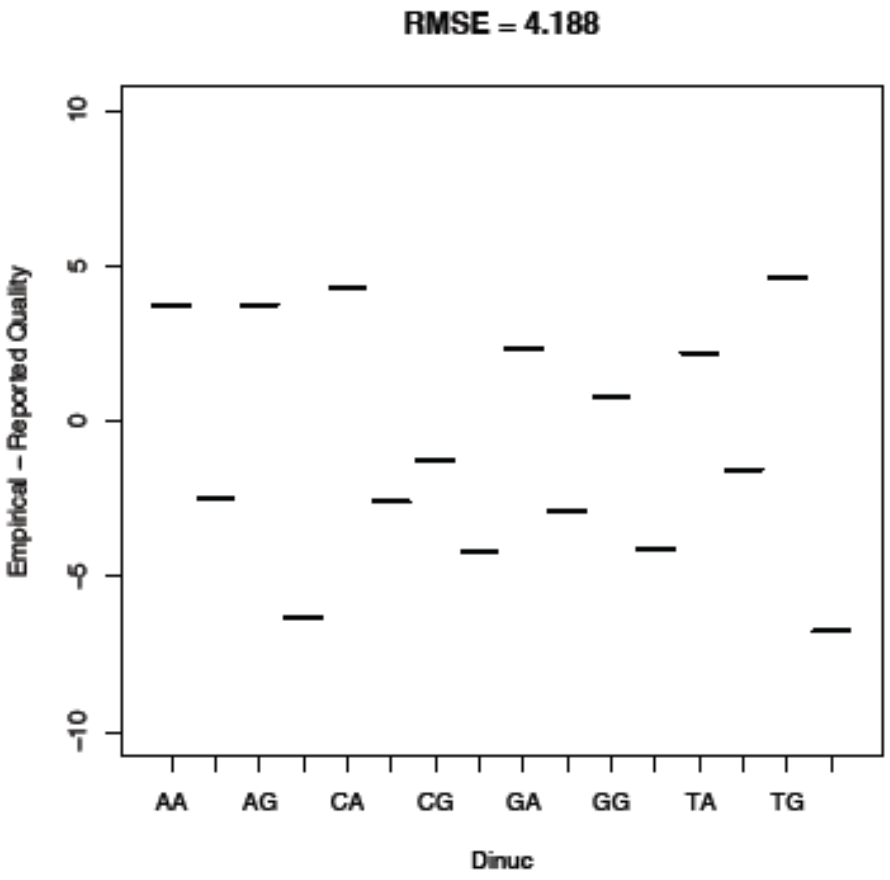
Before Recalibration



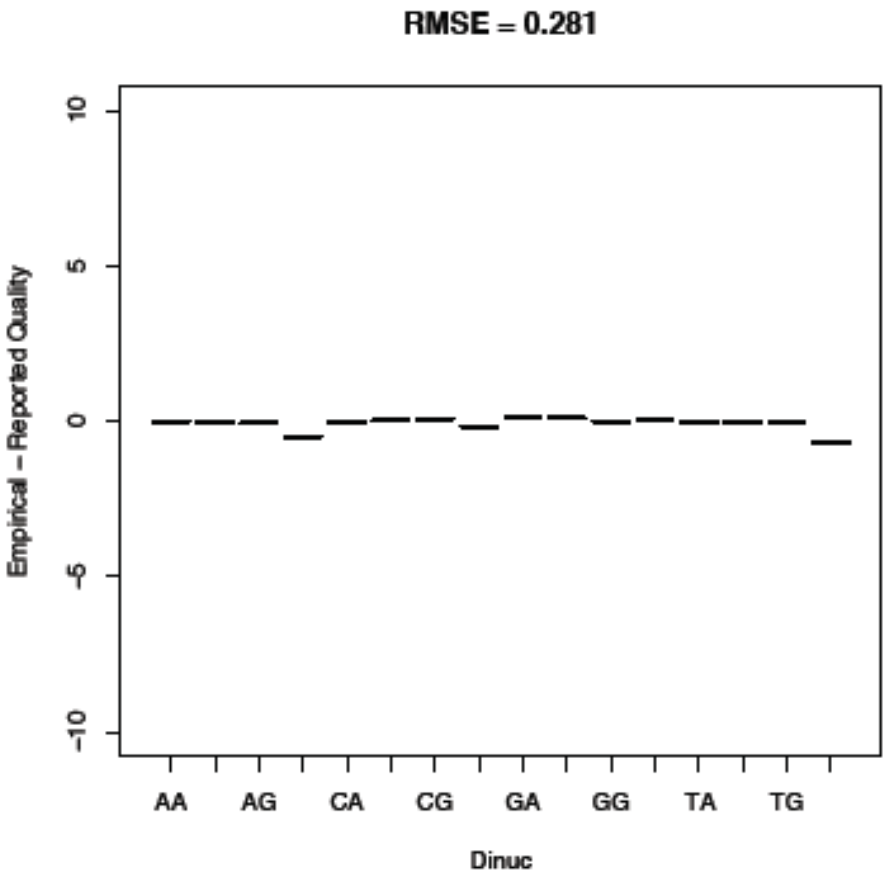
After Recalibration

Residual error by dinucleotide

SciLifeLab



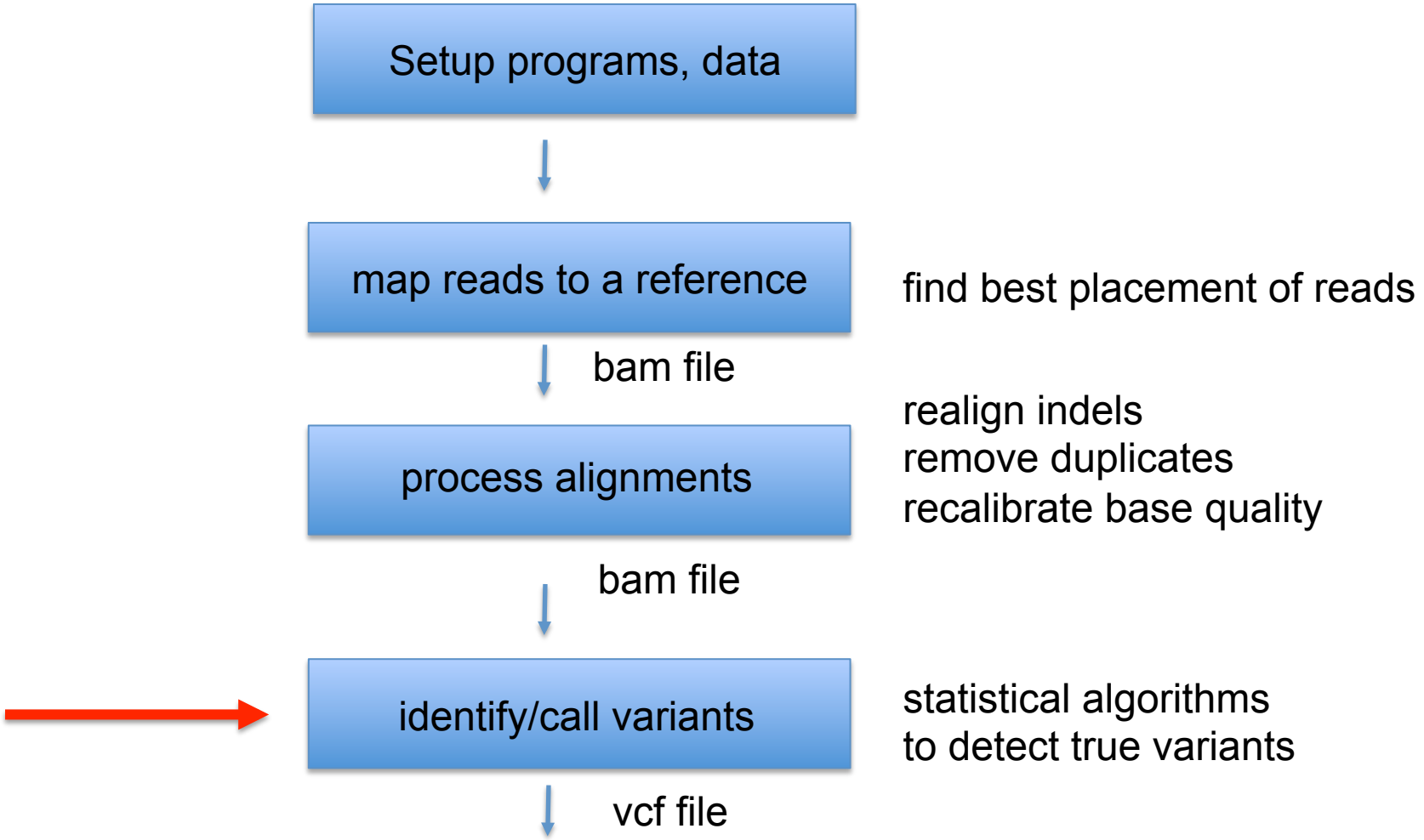
Before Recalibration



After Recalibration

Steps in resequencing analysis

SciLifeLab



Variant calling

SciLifeLab

simple pileup methods

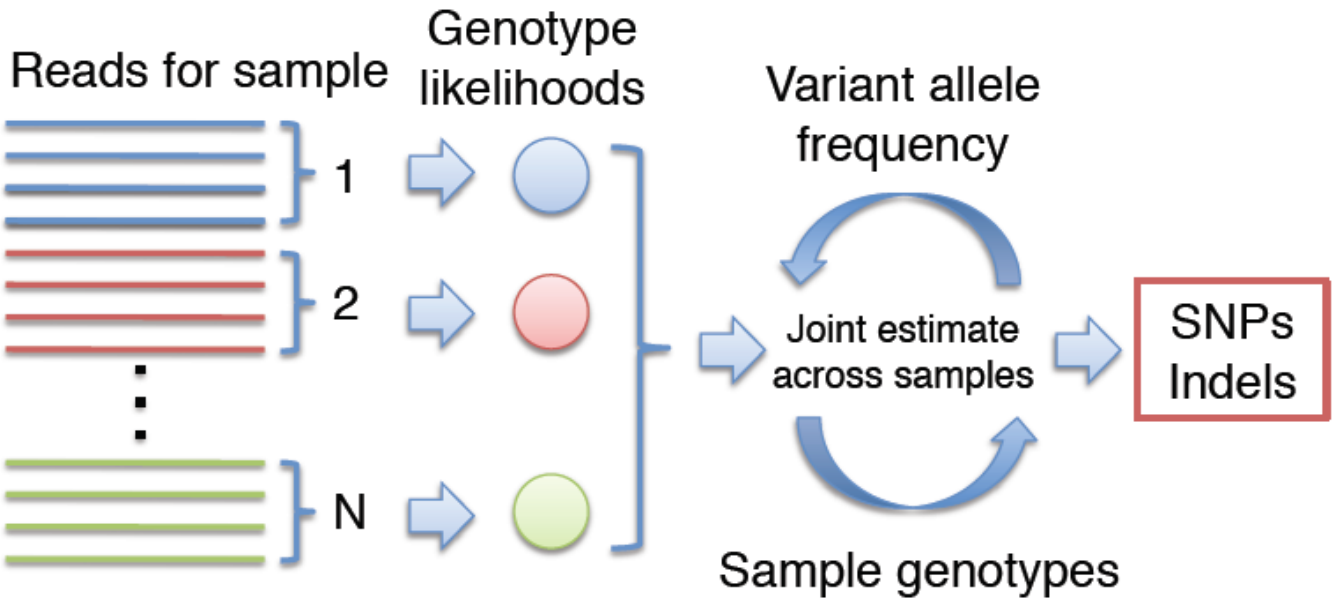
SciLifeLab

Reference: acacagatagacatagacatagacagatgag

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

Baysian population-based calling

SciLifeLab

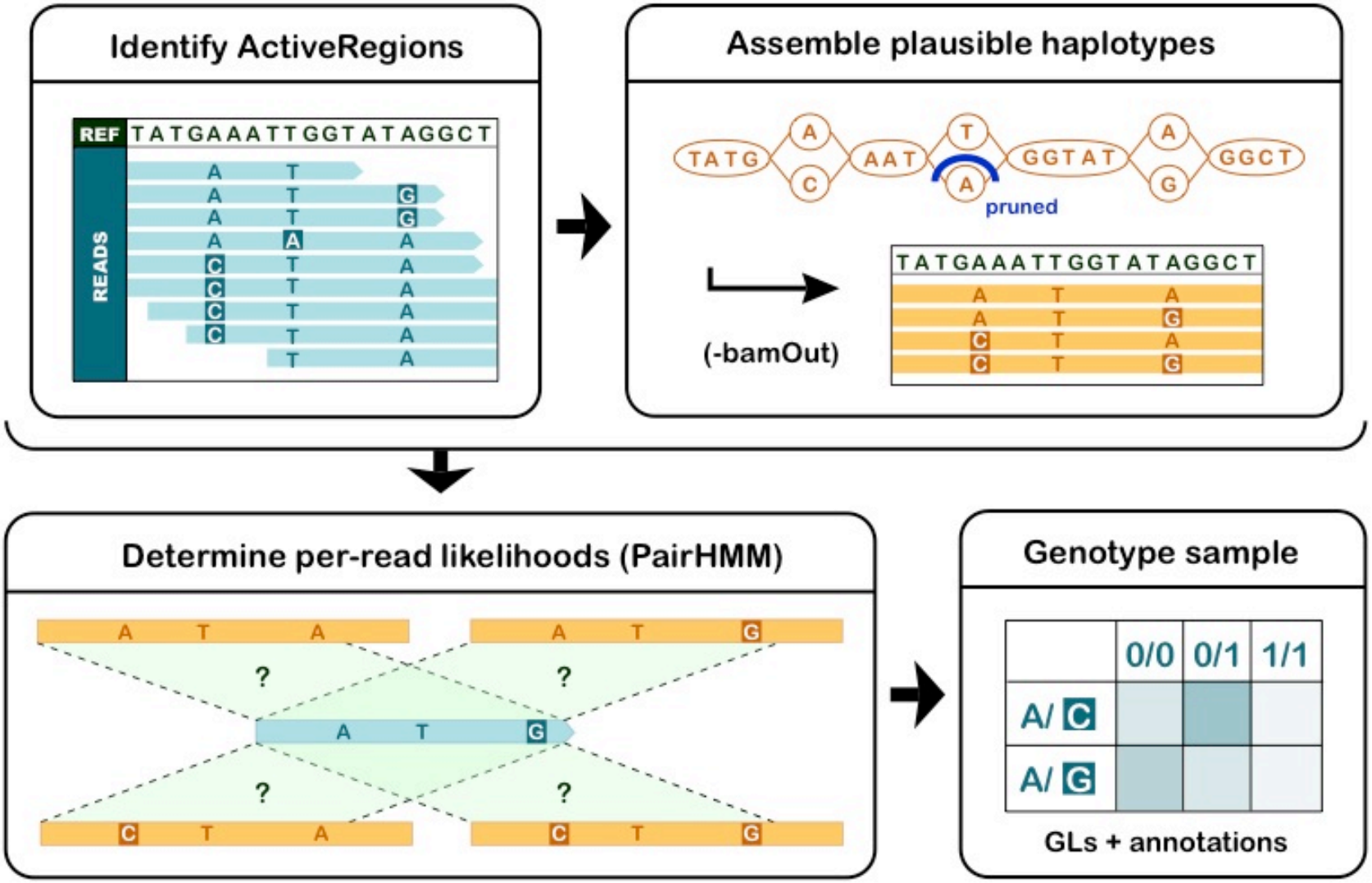


Simultaneous estimation of:

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

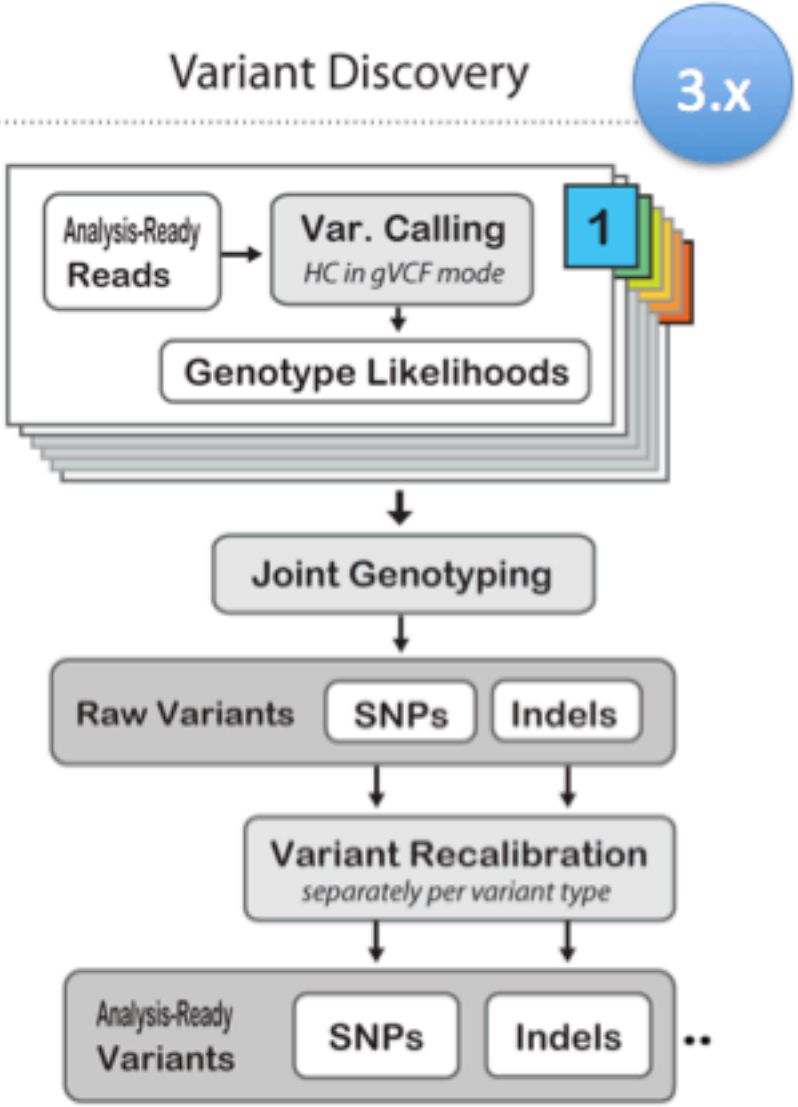
GATK haplotype caller

SciLifeLab



GATK best practice for cohorts

SciLifeLab



VCF format

SciLifeLab

```
##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

SciLifeLab

```
##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
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
```

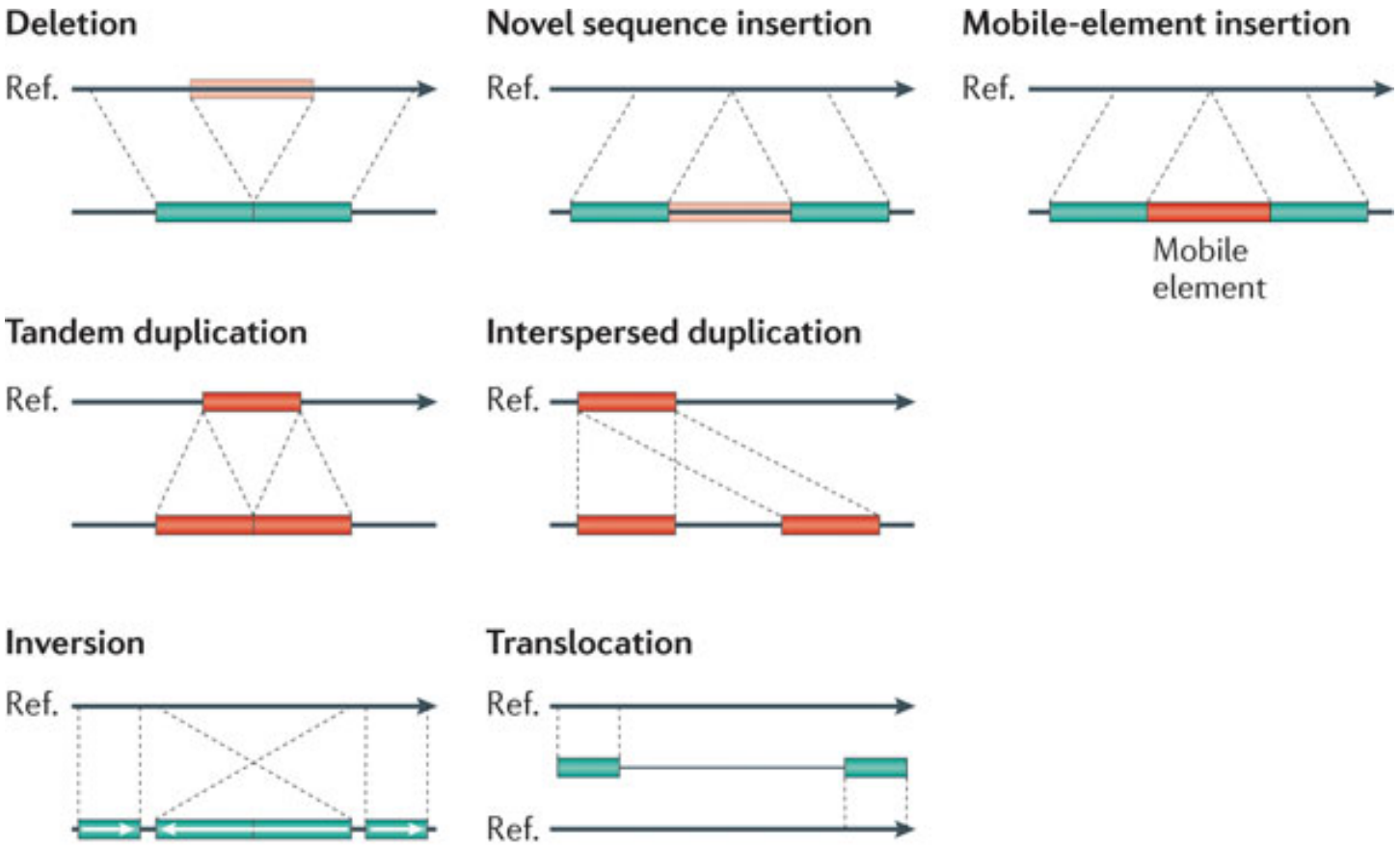
gVCF format

SciLifeLab

```
##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

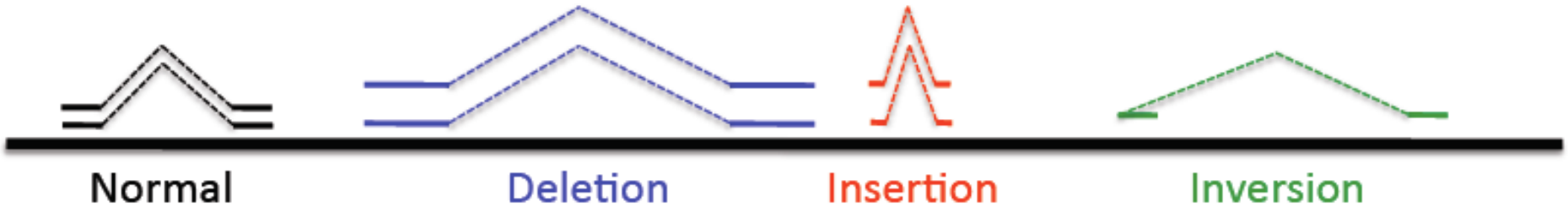
SciLifeLab

- Depth of coverage can be used to estimate copy number
- Samples may exhibit variation in depth indicative of polymorphic copy number variants
- How many copies of a duplication in the reference?
- How similar are the copies?
- Difficult to distinguish homozygotes and heterozygotes.

2) Paired end analysis

SciLifeLab

- 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

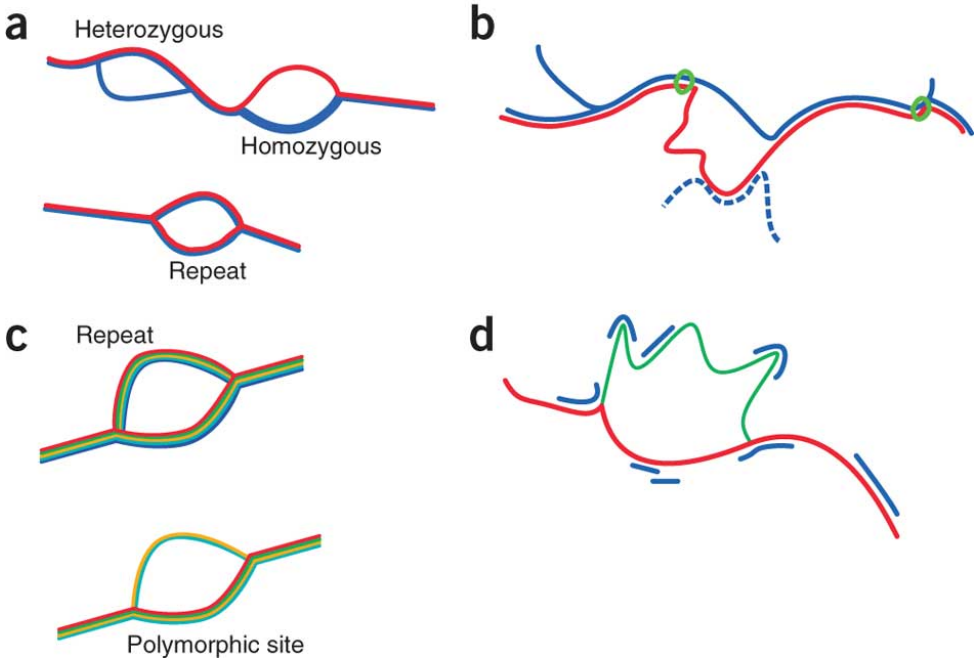
SciLifeLab

- 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

4) *De novo* assembly to identify structural variants

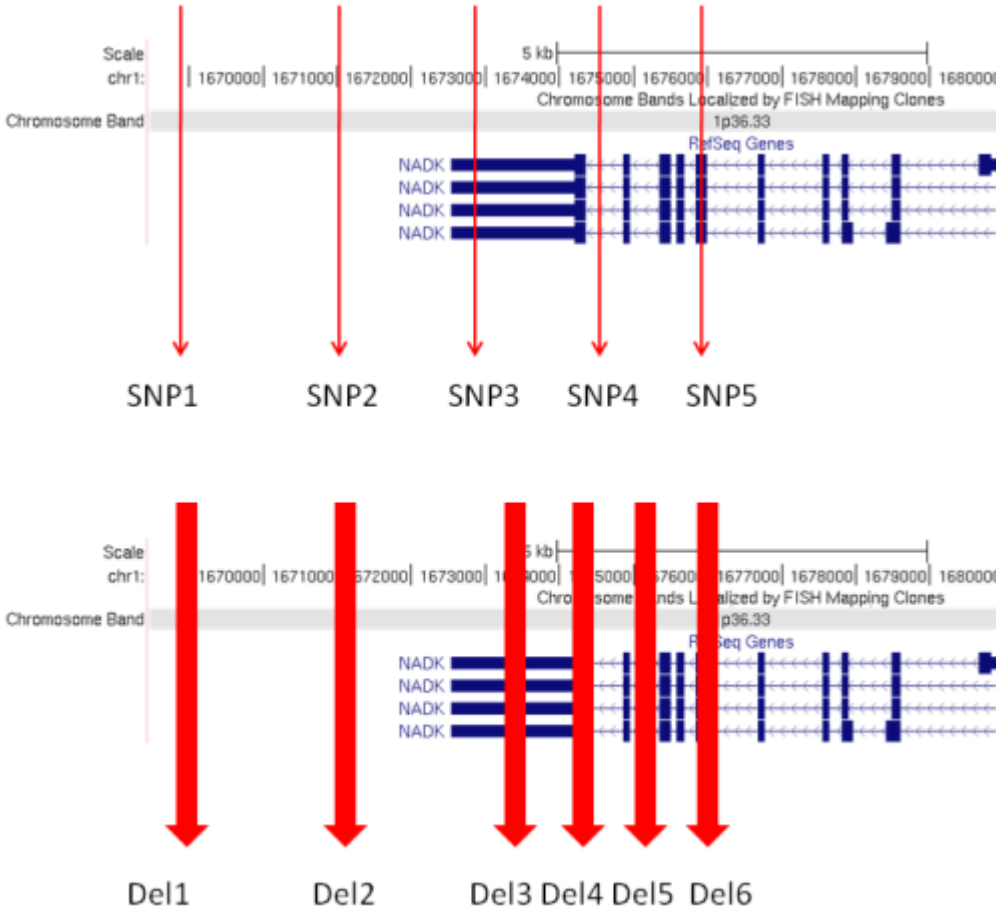
SciLifeLab

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



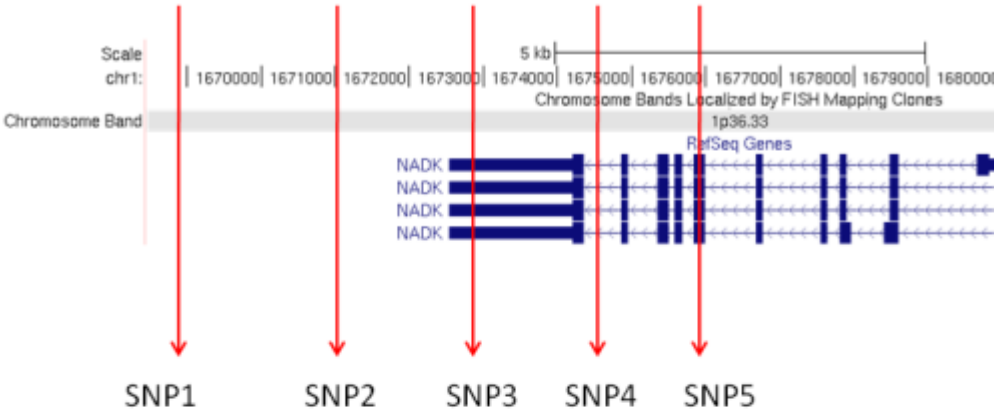
Annotation of variants

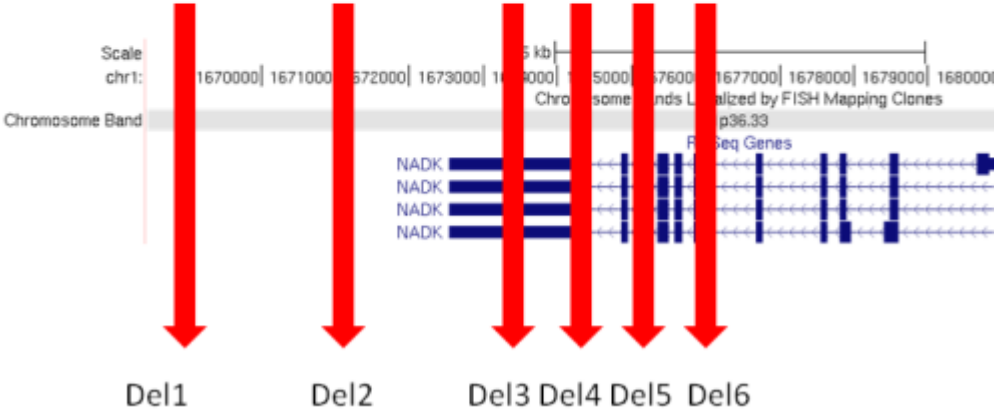
By comparing with existing annotation for the reference genome it is possible to gain information about localization and expected effect



Annotation of variants

By comparing with existing annotation for the reference genome it is possible to gain information about localization and expected effect





Most commonly used tools are Annovar and SNPEff

Downstream analysis

SciLifeLab

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

Overview of excercise

SciLifeLab

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 and programs

SciLifeLab

- 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

- Tasks: align a couple of samples to reference, process, reacalibration, call and filter variants

1) Access to data and programs

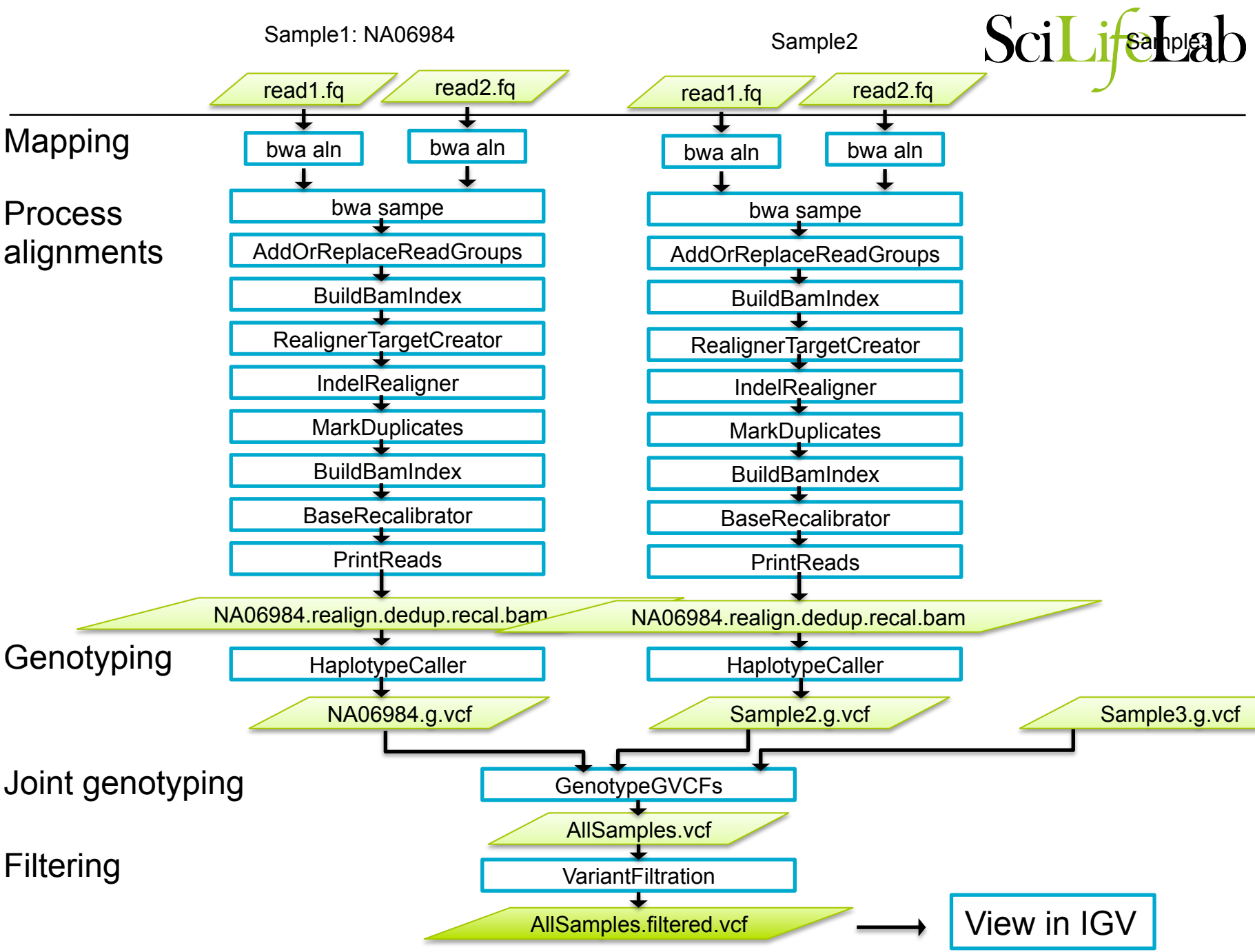
SciLifeLab

- 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/
```


Sample1: NA06984
Sample2
Sample3
SciLifeLab
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
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

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

SciLifeLab

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

SciLifeLab

- 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

SciLifeLab

- 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.

SciLifeLab

- 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

SciLifeLab

- 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

SciLifeLab

- 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

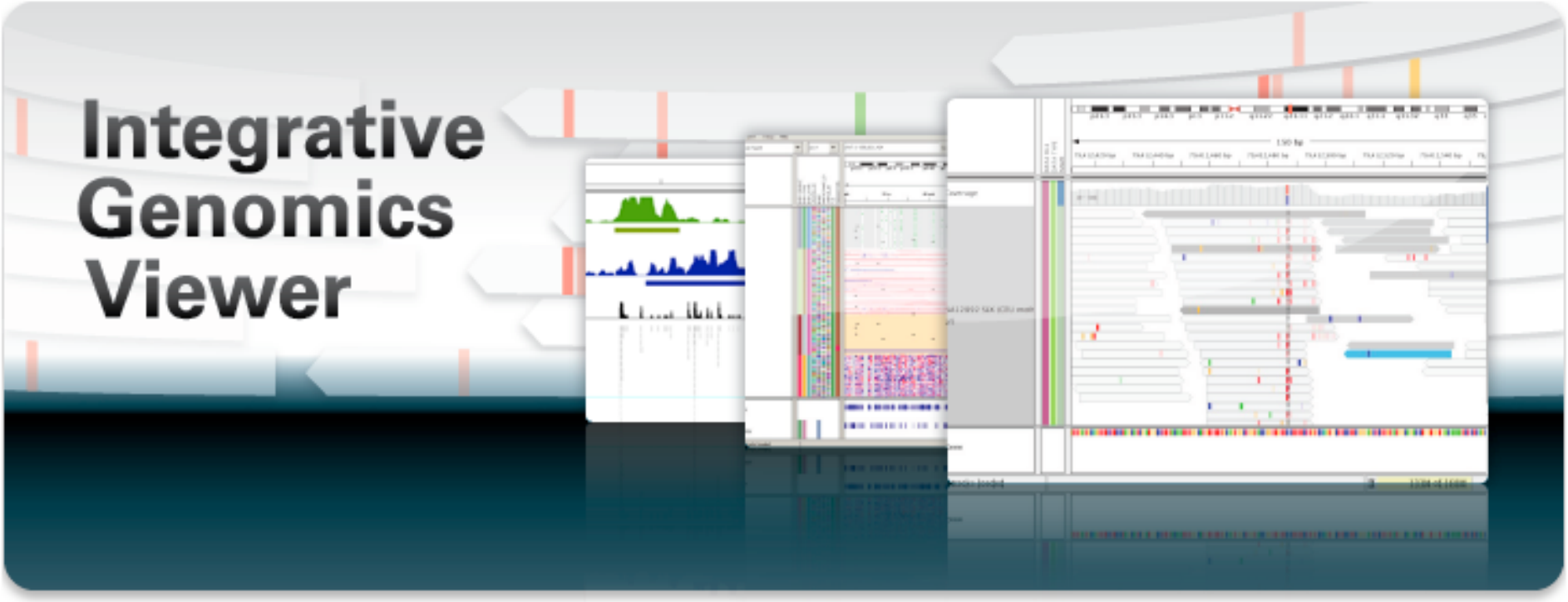
SciLifeLab

- variant filtering

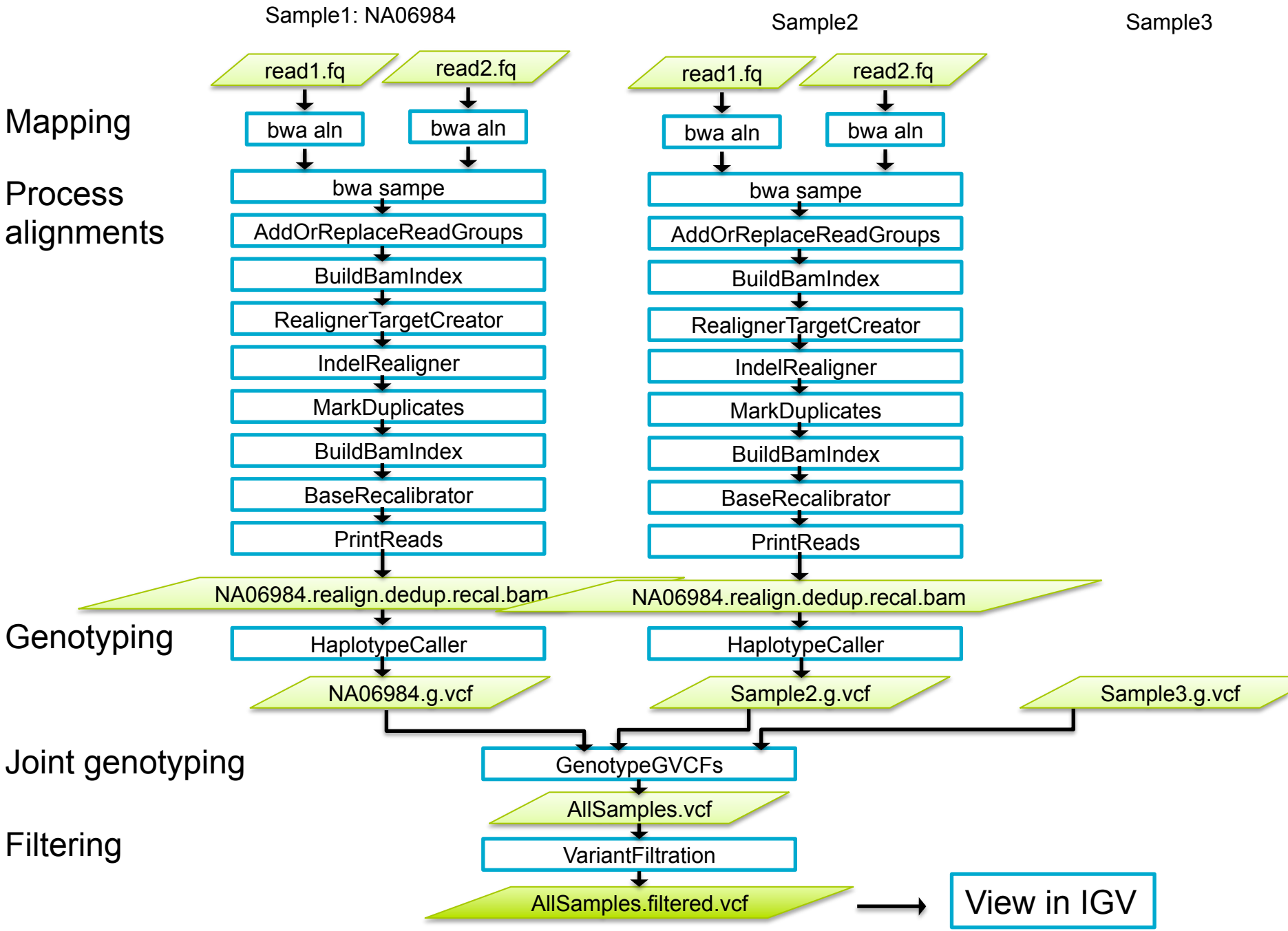
```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar
-T VariantFiltration
-R <reference>
-V <input vcf>
-o <output vcf>
--filterExpression "QD<2.0" --filterName QDfilter
--filterExpression "MQ<40.0" --filterName MQfilter
--filterExpression "FS>60.0" --filterName FSfilter
--filterExpression "HaplotypeScore>13.0"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

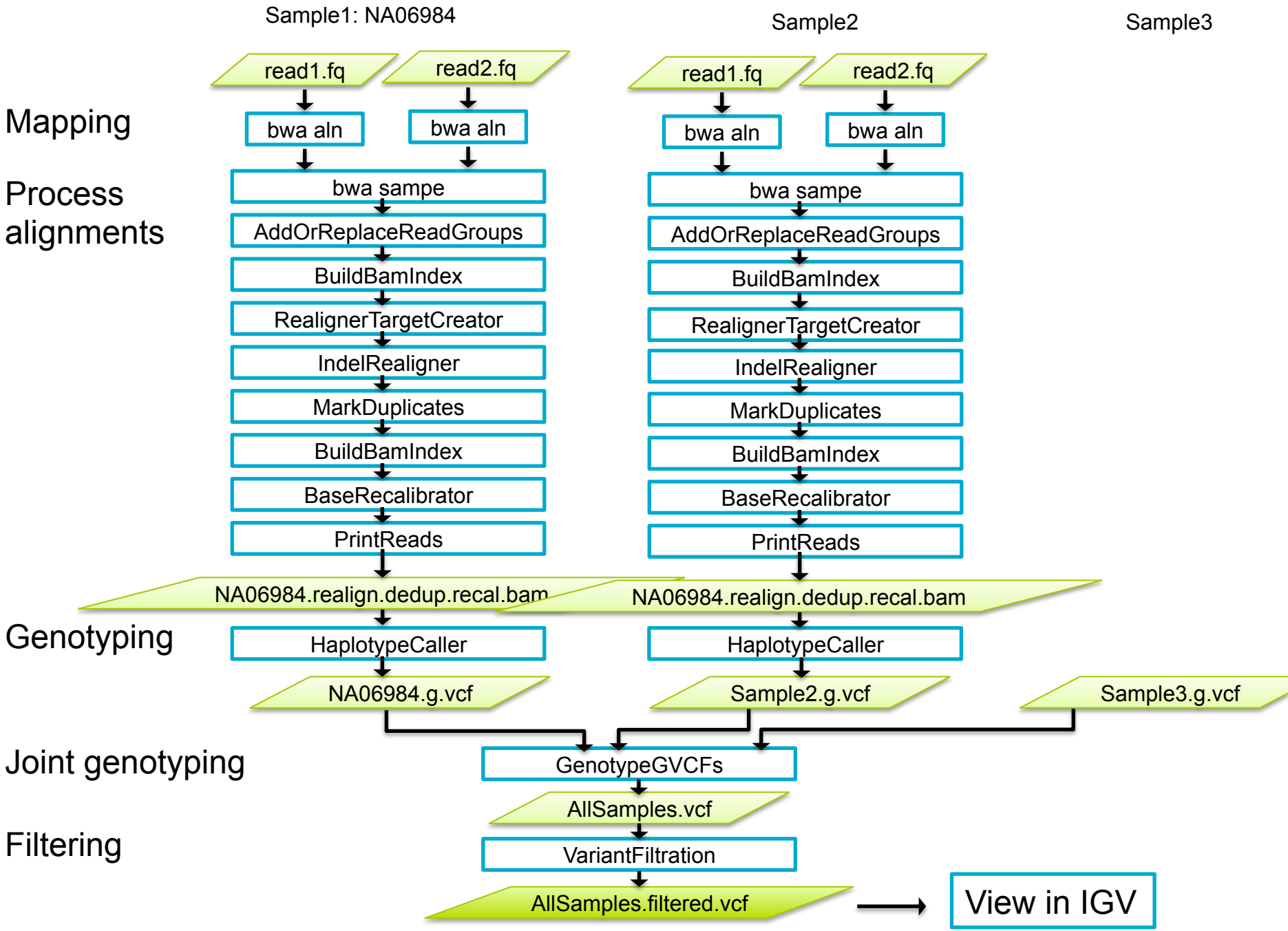
7) Viewing data with IGV

SciLifeLab



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

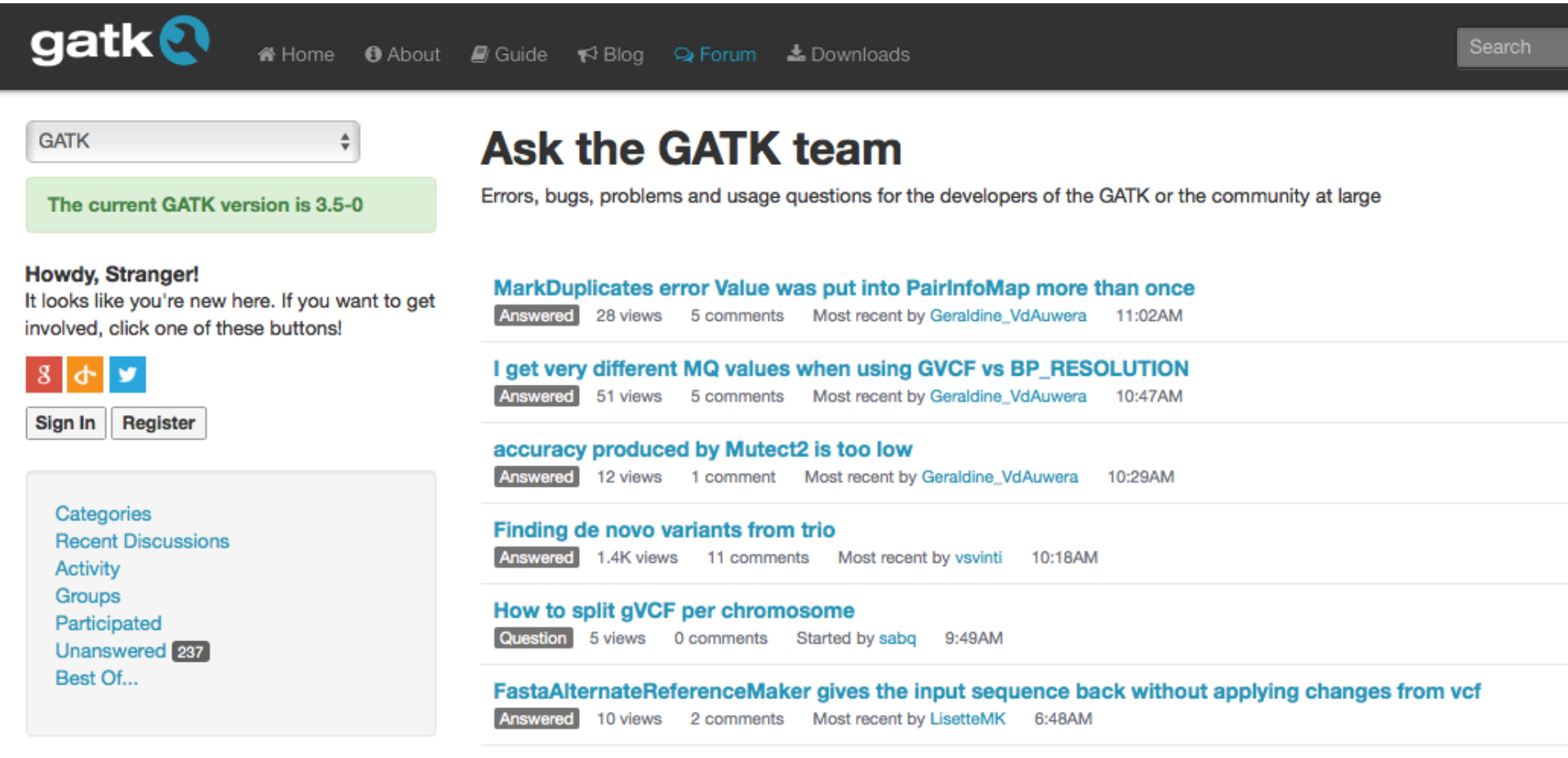

Sample1: NA06984
Sample2
Sample3
Mapping
Process alignments
Genotyping
Joint genotyping
Filtering
read1.fq
read2.fq
bwa aln
bwa aln
bwa sampe
AddOrReplaceReadGroups
BuildBamIndex
RealignerTargetCreator
IndelRealigner
MarkDuplicates
BuildBamIndex
BaseRecalibrator
PrintReads
NA06984.realign.dedup.recal.bam
HaplotypeCaller
NA06984.g.vcf
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.g.vcf
GenotypeGVCFs
AllSamples.vcf
VariantFiltration
AllSamples.filtered.vcf
View in IGV


Sample1: NA06984
Sample2
Sample3
Mapping
Process alignments
Genotyping
Joint genotyping
Filtering
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
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.g.vcf
GenotypeGVCFs
AllSamples.vcf
VariantFiltration
AllSamples.filtered.vcf
View in 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



X) Extra

SciLifeLab

Extra 1: View data in UCSC-browser

Extra 2: Select subset with BEDTools

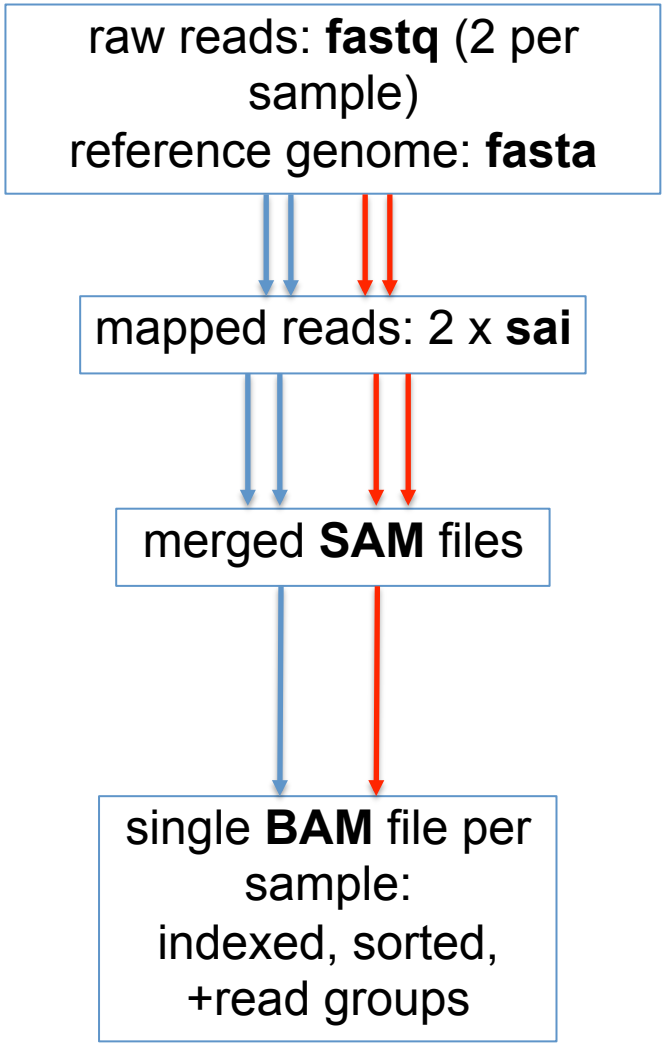
Extra 3: Annotate variants with annovar

Extra 4: Make a script to run pipeline

pipeline (1)

SciLifeLab

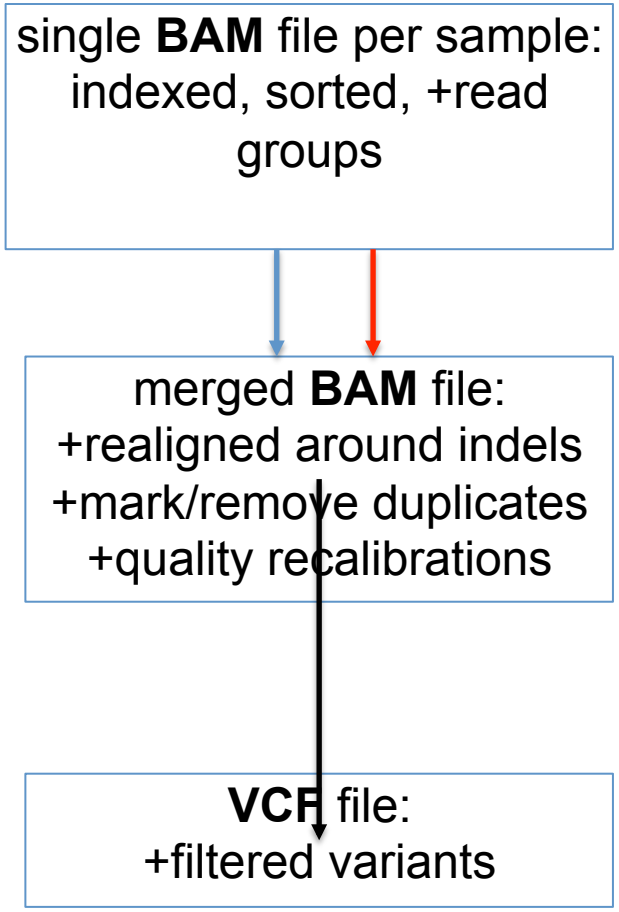
2. Mapping
   - `bwa index`
   - `samtools faidx`
   - `bwa aln`
3. Merging alignments
   - `bwa sampe`
4. Creating BAM files
   - `picard AddOrReplaceReadGroups`
   - `picard BuildBamIndex`

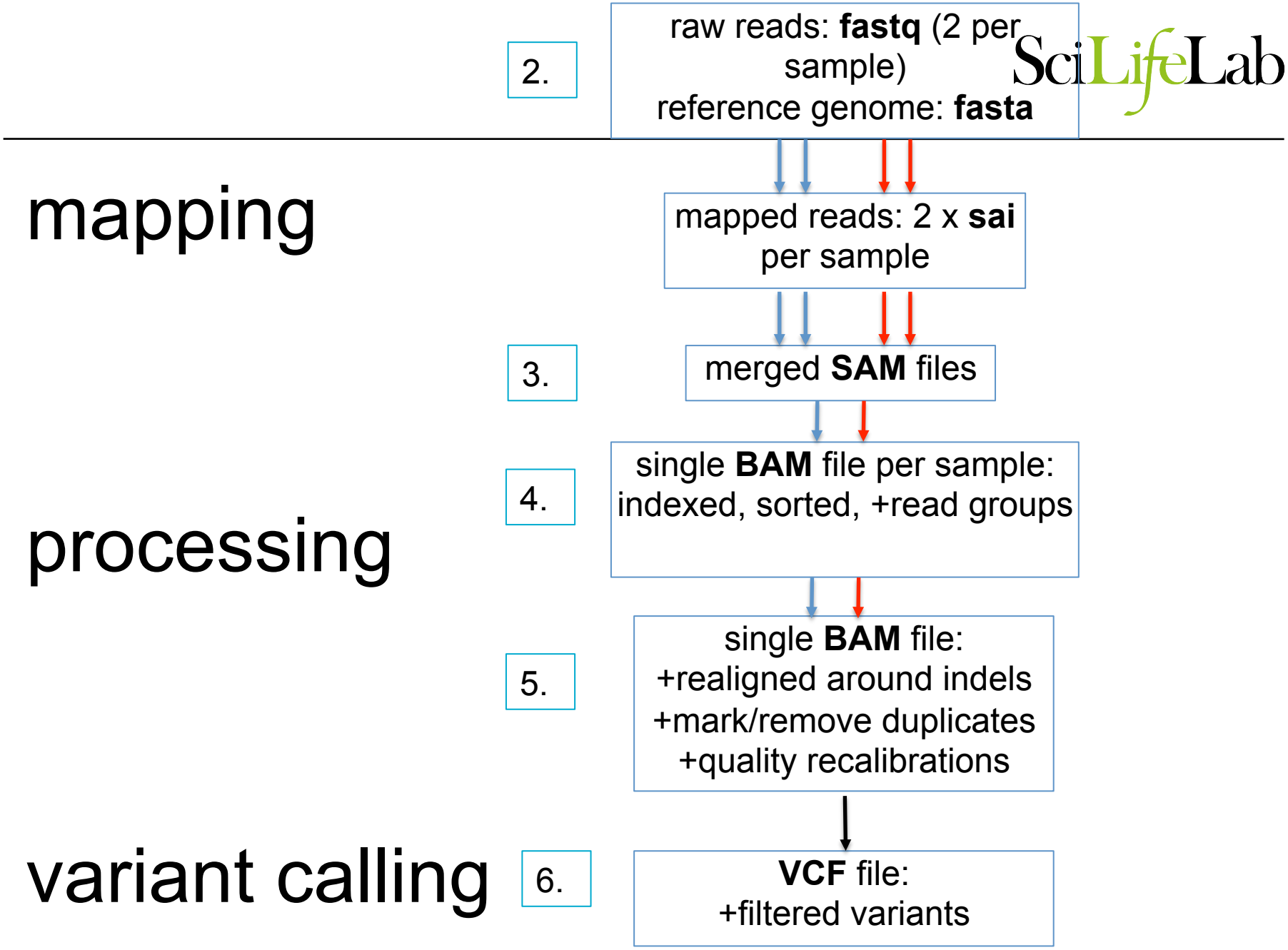


pipeline (2)

SciLifeLab

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)




2.
raw reads: **fastq** (2 per sample)
reference genome: **fasta**
mapping
mapped reads: 2 x **sai** per sample
3.
merged **SAM** files
processing
4.
single **BAM** file per sample: indexed, sorted, +read groups
5.
single **BAM** file:
+realigned around indels
+mark/remove duplicates
+quality recalibrations
variant calling
6.
VCF file:
+filtered variants

Naming conventions

SciLifeLab

Initial file name according to information about the content

NA06984.ILLUMINA.low_coverage.17q

For each step of the pipeline, 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

…