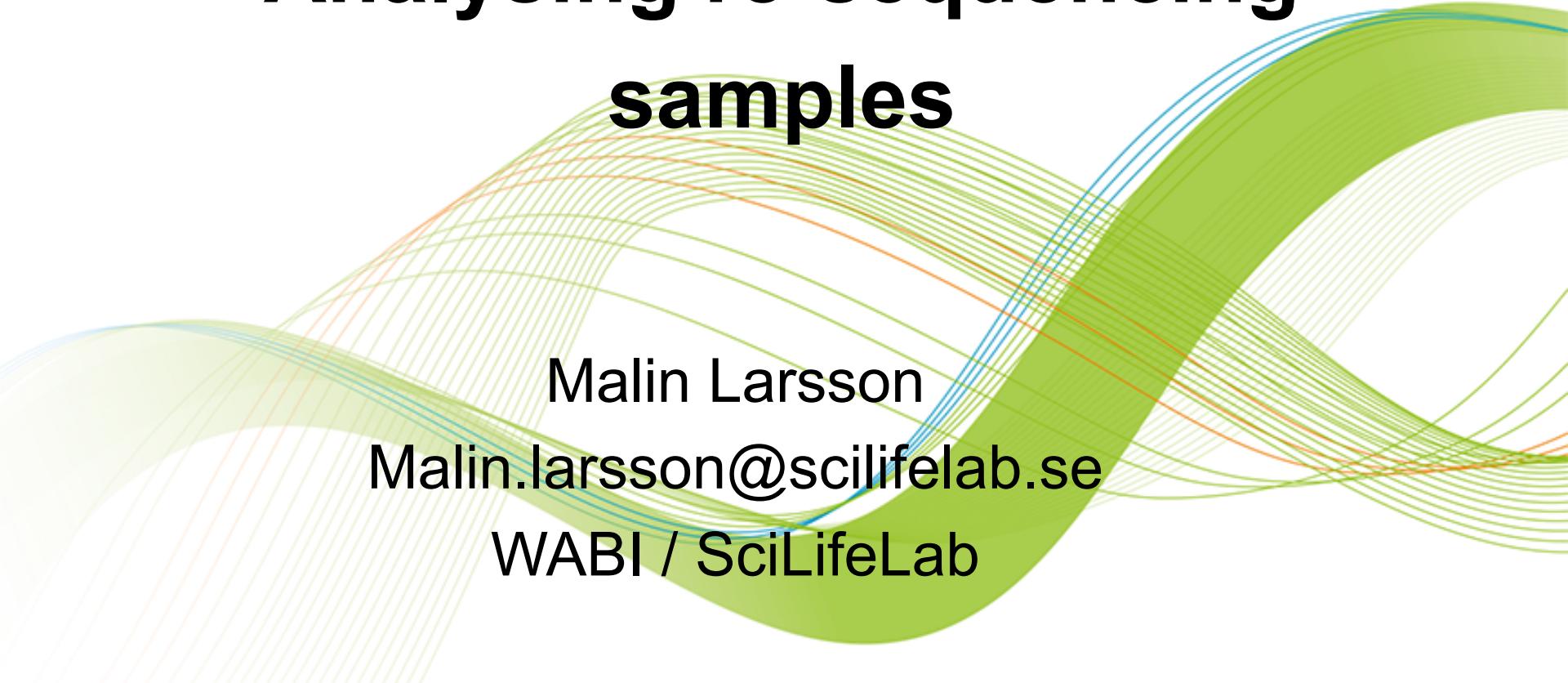




Analysing re-sequencing samples

The background of the slide features a series of overlapping, wavy lines in shades of green, blue, and orange, creating a sense of depth and motion. These lines are concentrated in the lower half of the slide, forming a base for the text.

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

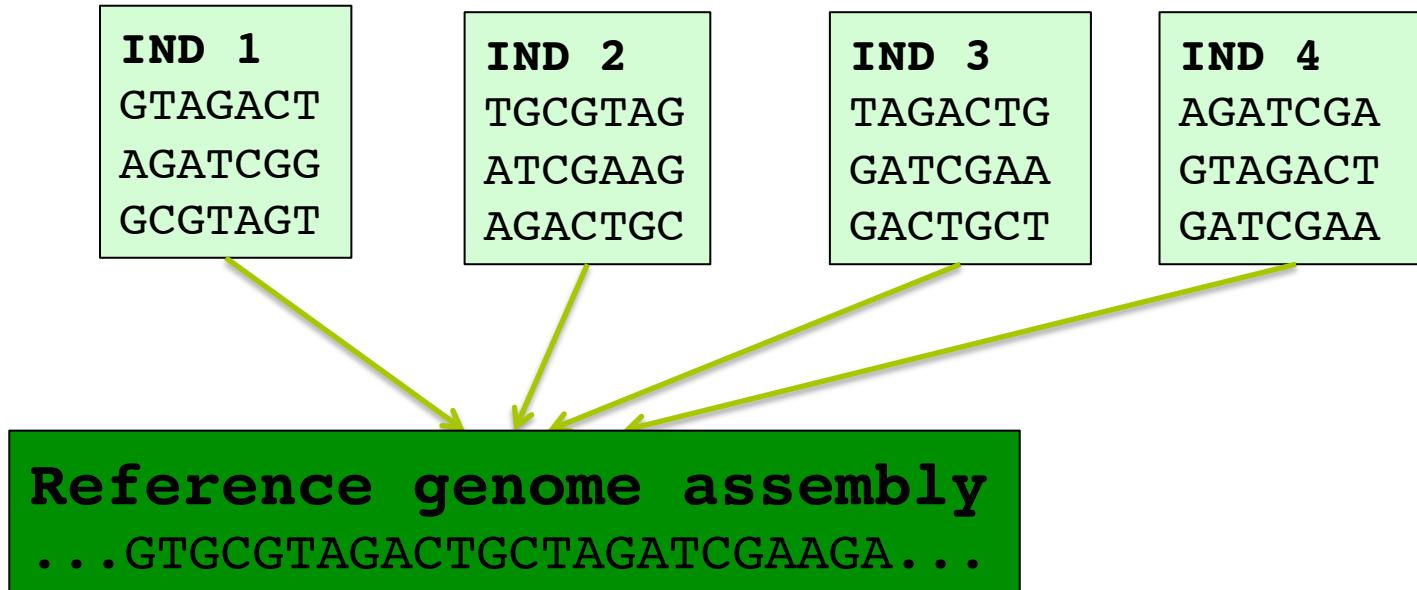
IND 4

AGATCGA
GTAGACT
GATCGAA

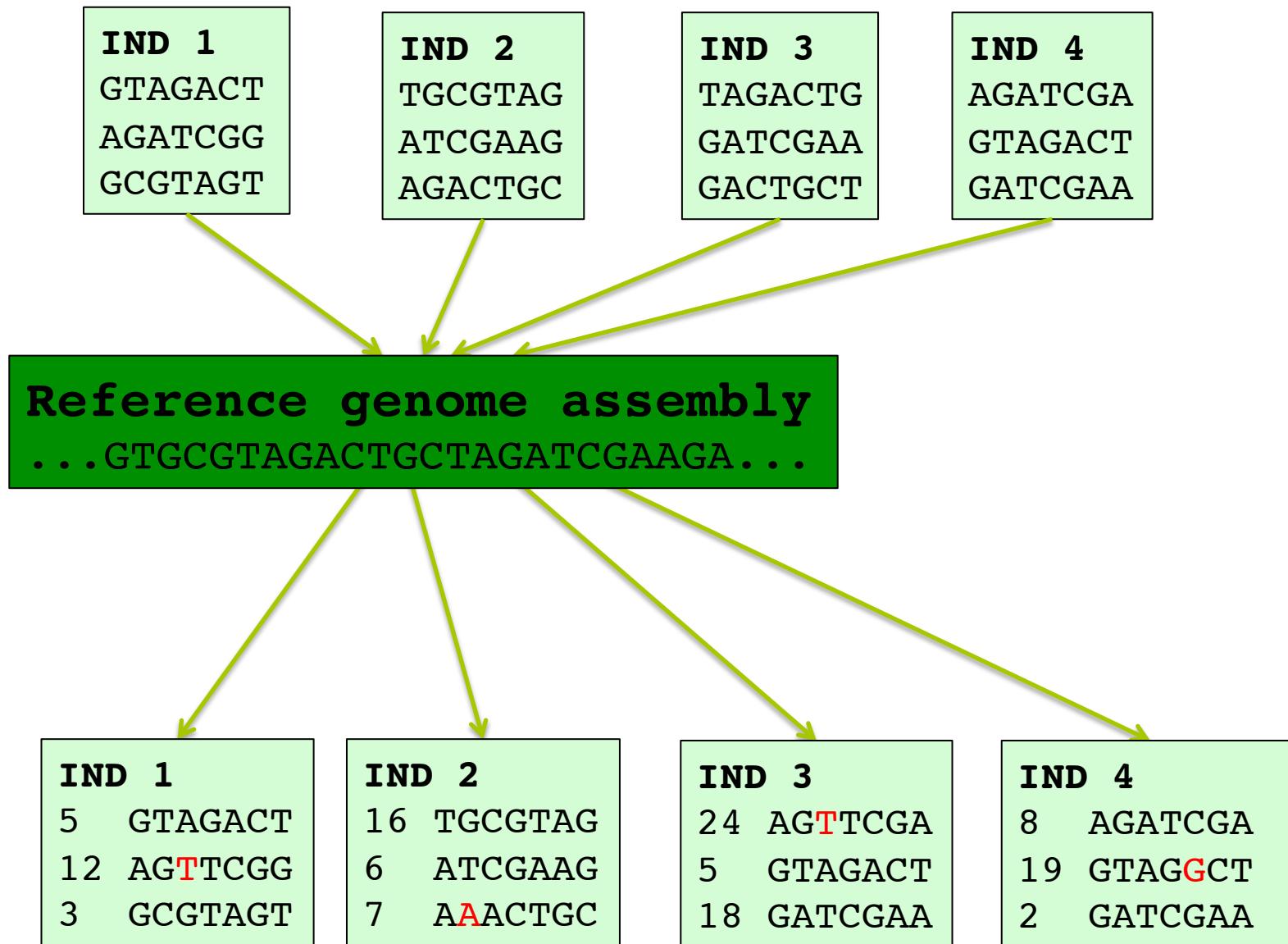
Reference genome assembly

...GTGCGTAGACTGCTAGATCGAAGA...

Re-sequencing



Re-sequencing



Rare variants in human

SciLifeLab

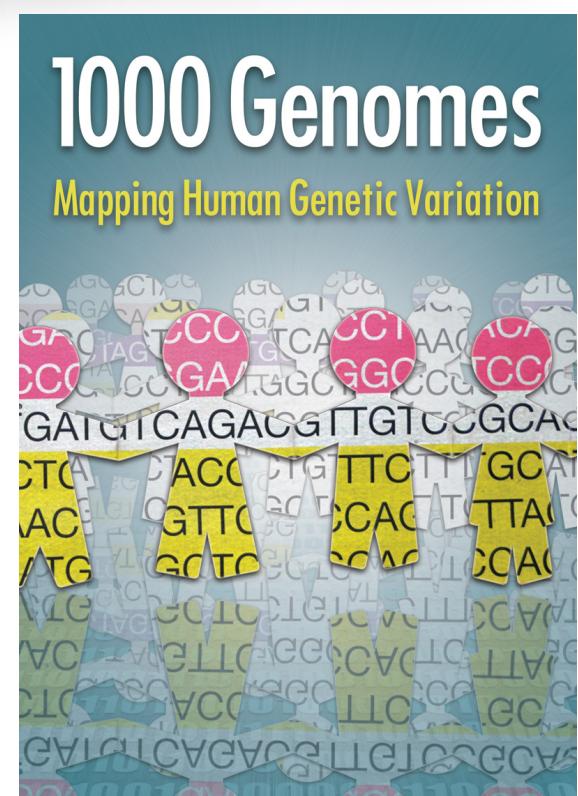


deCODE genetics

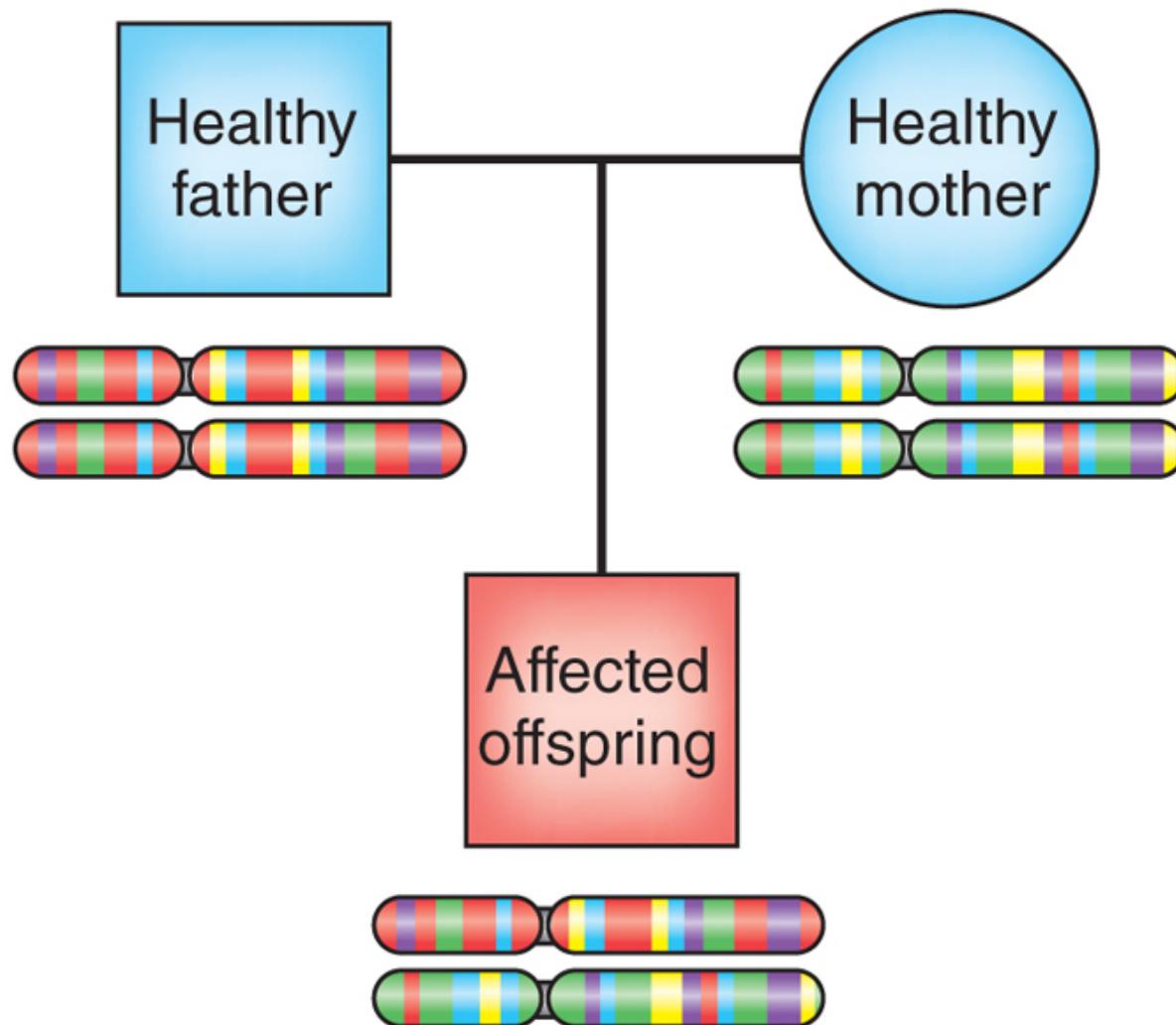


UK
10K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE



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



Population genetics – speciation, adaptive evolution

Darwin Finches

b

1 *G. magnirostris_G*



2 *G. difficilis_W*



3 *G. difficilis_P*



4 *T. bicolor_B*



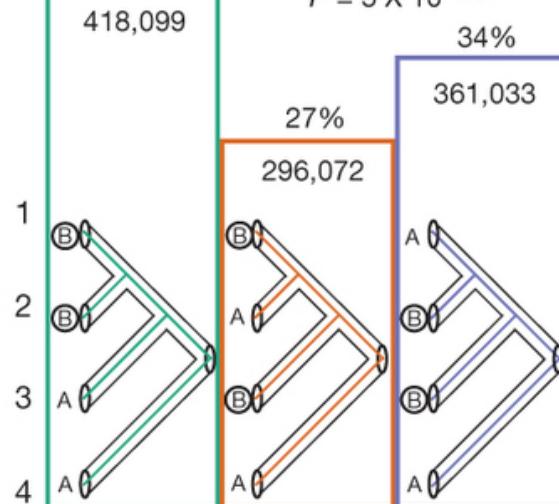
39%

418,099

$$D = 0.098$$

$$P = 5 \times 10^{-113}$$

34%



Population genetics – speciation, adaptive evolution

Darwin Finches

b
 1 *G. magnirostris*_
 2 *G. difficilis*_V
 3 *G. difficilis*_
 4 *T. bicolor*_

39%

418,099

27%

296,072

P =

<0.0001

0.0001

0.001

0.01

0.1

1

1

2

3

4

Heliconius Butterflies

support
 strap support
 trap support

Melpomene_rosina.2097

Melpomene_rosina.2071

Melpomene_melpomene.14671

Melpomene_melpomene.114671

Melpomene_melpomene.CM7

Melpomene_melpomene.CM8

Melpomene_melpomene.CM9

Melpomene_melpomene.CM1

Melpomene_melpomene.CM3

Melpomene_melpomene.CM2

Melpomene_melpomene.8228

Melpomene_melpomene.8229

Melpomene_aglaope.09-247

Melpomene_aglaope.09-128

Melpomene_aglaope.09-119

Melpomene_aglaope.09-107

Melpomene_aglaope.09-102

Melpomene_amaryllis.09-76

Melpomene_amaryllis.09-8

Melpomene_amaryllis.09-334

Melpomene_amaryllis.09-7

Melpomene_amaryllis.09-1

Melpomene_ecuadorensis.9111

Melpomene_ecuadorensis.9112

Timareta_sep.nov.09-232

Timareta_sep.nov.09-233

Timareta_sep.nov.09-231

Timareta_sep.nov.09-234

Timareta_sep.nov.09-68

Timareta_timareta.9179

Timareta_timareta.9180

Heurippa.CH11

Heurippa.CH14

Heurippa.CH7

Heurippa.CH8

Heurippa.CH9

Heurippa.CH12

Cydno_cordula.M2253

Cydno_cordula.M2166

Cydno_cordula.M2255

Cydno_cordula.M2186

Cydno_cordula.M2259

Cydno_cordula.M2159

Cydno_weymeri.204

Cydno_weymeri.221

Cydno_cyndies.217

Cydno_cyndies.216

Cydno_chioneus.8265

Cydno_chioneus.2440

H. hecale felix (Peru)

H. melpomene
rosina (Panama)

H. melpomene
melpomene (Panama)

H. melpomene
melpomene
 (Colombia)

H. melpomene
melpomene
 (French Guiana)

H. melpomene
aglaope
 (Peru)

H. melpomene
amarillys (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 cyndies
 (Colombia)

H. cydno chioneus
 (Panama)

H. hecale felix (Peru)

H. melpomene

H. timareta

H. heurippa

H. cydno cordula

H. cydno weymeri

H. cydno cyndies

H. cydno chioneus

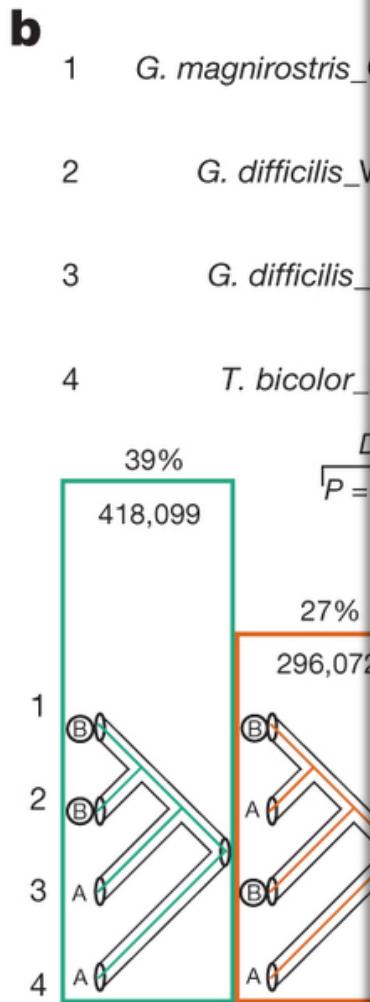
H. hecale felix

silvaniforms

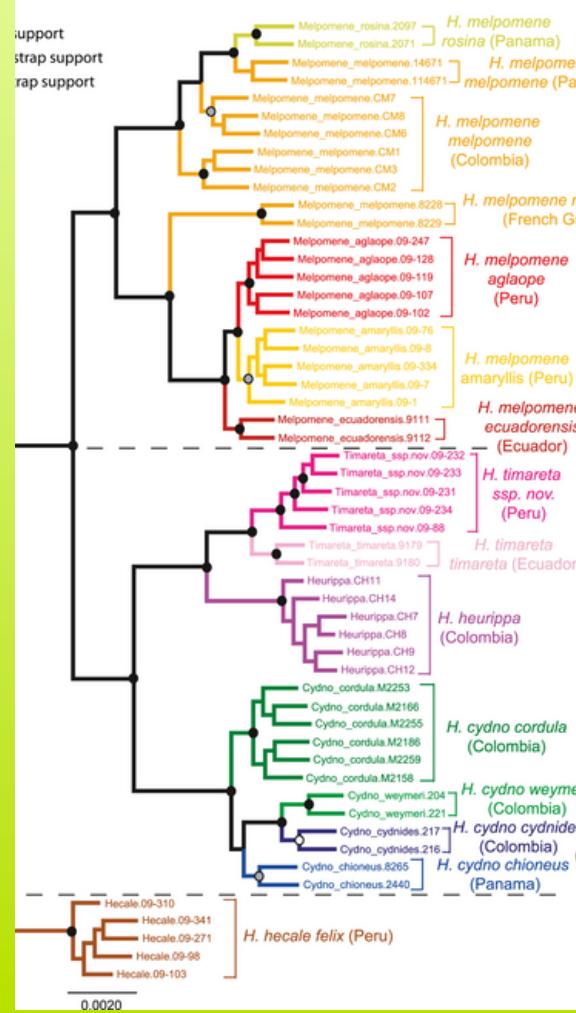
0.0020

Population genetics – speciation, adaptive evolution

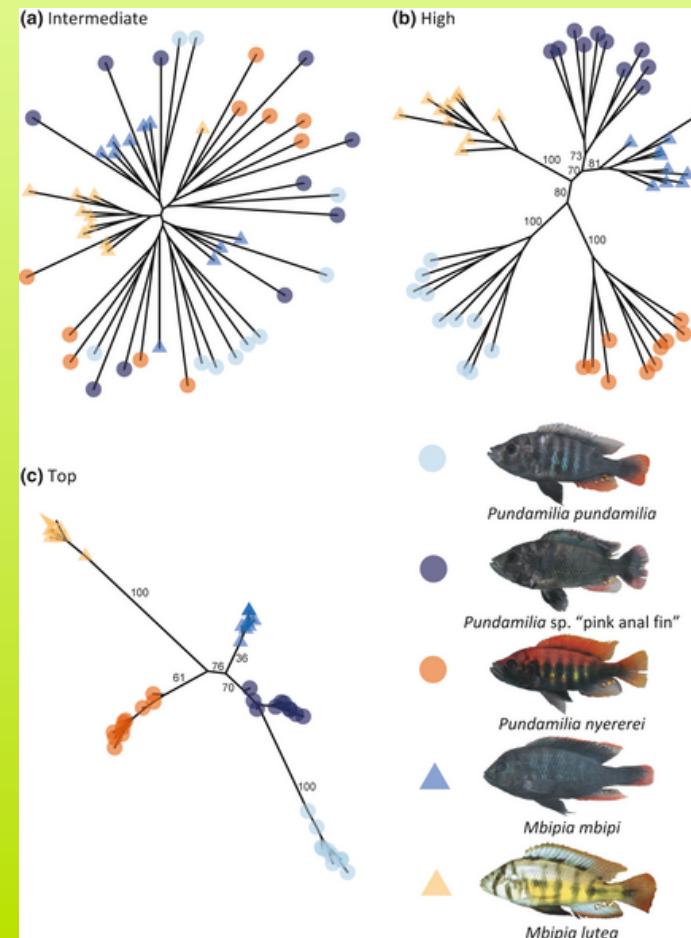
Darwin Finches



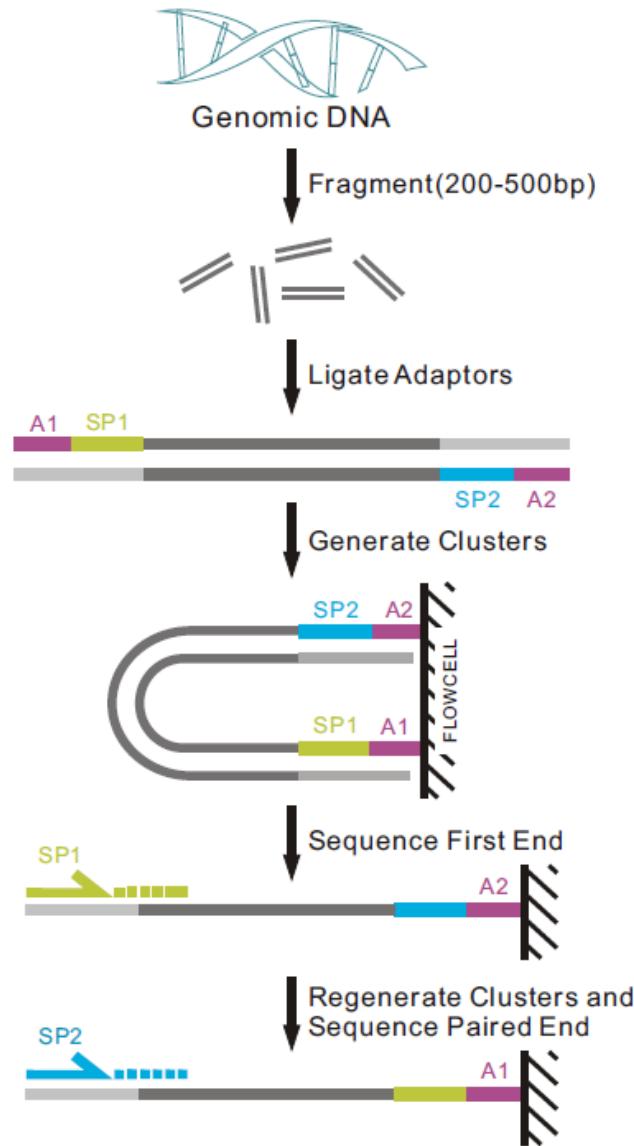
Heliconius Butterflies



Lake Victoria cichlid 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  
ATATGGAACGTTGTGGTCTGAAAGAAGATGT  
+  
B@CFFFFFHHHHGJJJJJJJJFHHIIIIJJ  
JIHGIIJJJJJIJIIJJJJIIJJJJIIIEIHHIJ  
HGHHHHHDFFFEDDDDDCDDDCDDDDDDCDC
```

ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:  
5:1101:1160:2197 2:N:0:ATCACG  
CTTCGTCCACTTCATTATTCCCTTCATACATG  
CTCTCCGGTTAGGGTACTCTGACCTGGCCTT  
TTTTCAAGACGTCCCTGACTTGATCTGAAACG  
+  
CCCFFFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJ  
JJJJJJJJJJJIJGIJHBGHIIIIJJJJJJJJJJ  
JJJHFFFFFFDDDDDDDDDDDDDDEDCCDDDD
```

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

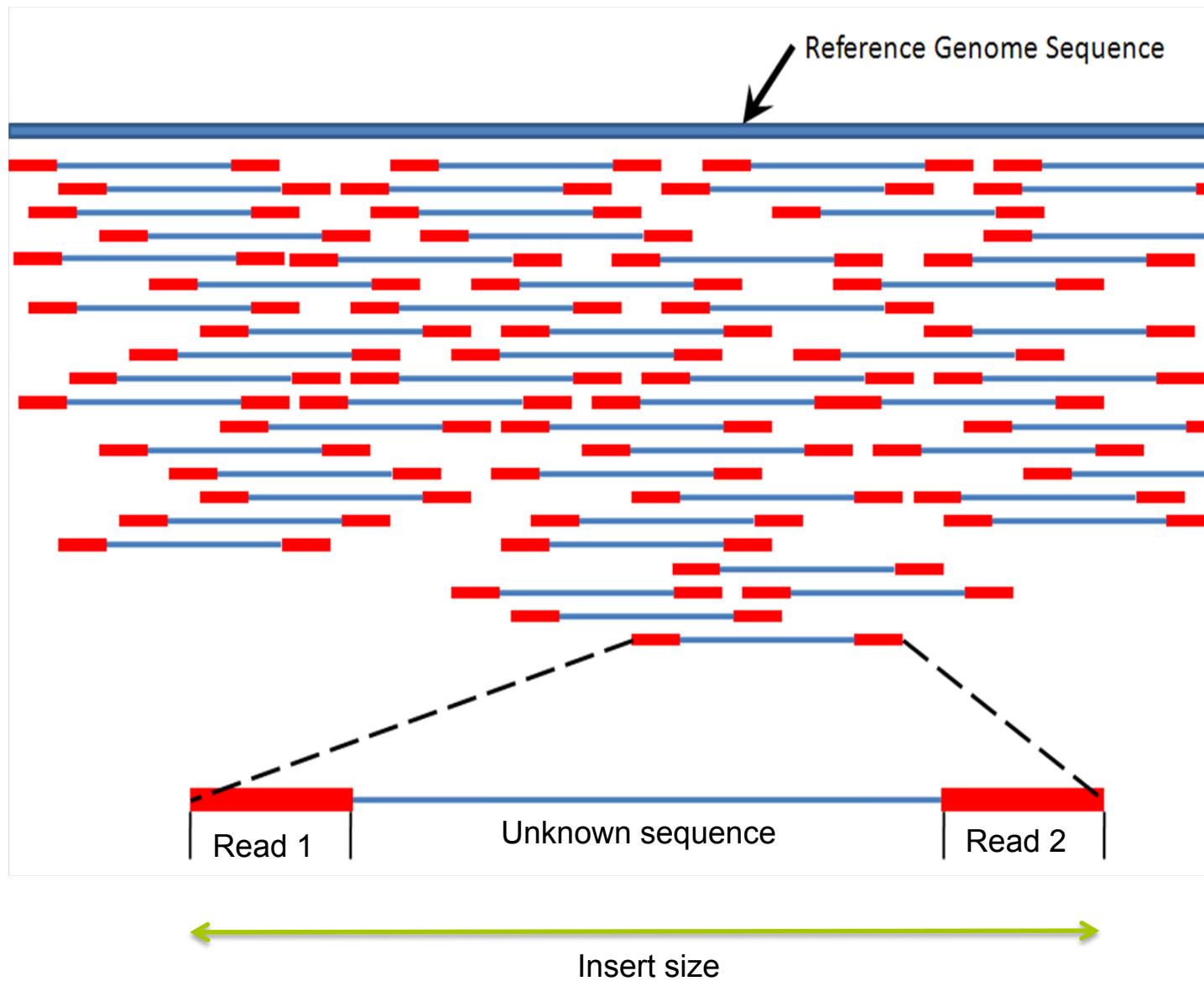
ID_R1_001.fastq

```
@HISEQ:100:C3MG8ACXX:  
5:1101:1160:2197 1:N:0:ATCACG  
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG  
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG  
ATATGGAACGTTGTGGTCTGAAAGAAGATGT  
+  
B@CFFFFFHGGJJJJJJJFHHIIIIJJ  
JIHIIJJJJIIJJJJJJIIIEIHHIJ  
HGHHHHHDFFFEDDDDCDDDCDDDDDCDC
```

ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:  
5:1101:1160:2197 2:N:0:ATCACG  
CTTCGTCCACTTCATTATTCCTTCATACATG  
CTCTCCGGTTAGGGTACTCTGACCTGGCCTT  
TTTTCAAGACGTCCCTGACTTGATCTGAAACG  
+  
CCCFFFFFHGGJJJJJJJJJJJJJJJJJJJJJJJJ  
JJJJJJJIJGIJHGHHIIIIJJJJJJJJJJJJJJ
```

Paired end sequencing



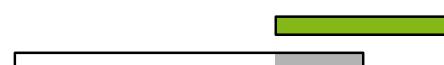
Adapter trimming

Module load cutadapt

3' Adapter



or



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

5' Adapter



or



Anchored 5' adapter



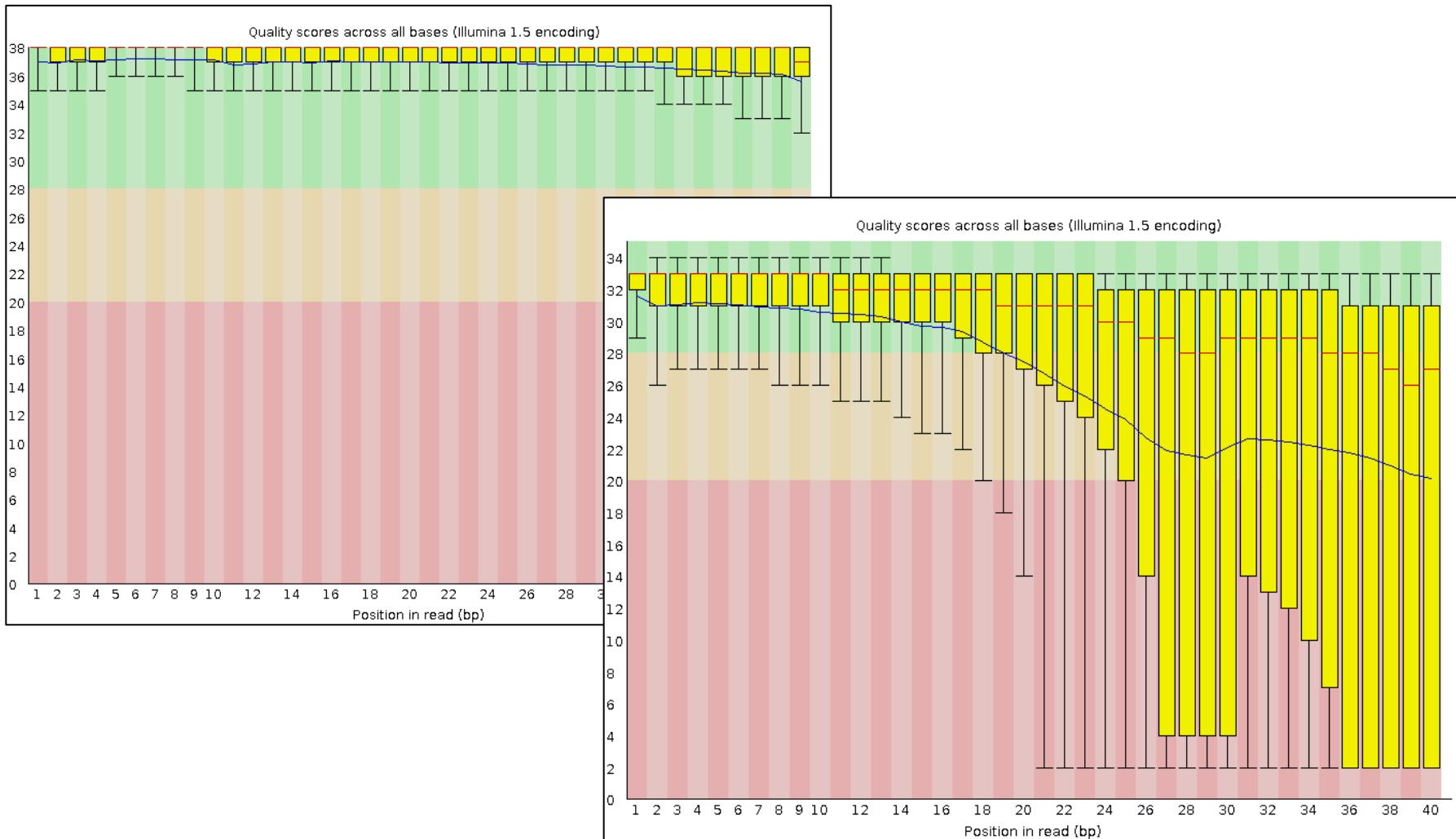
Read

Adapter

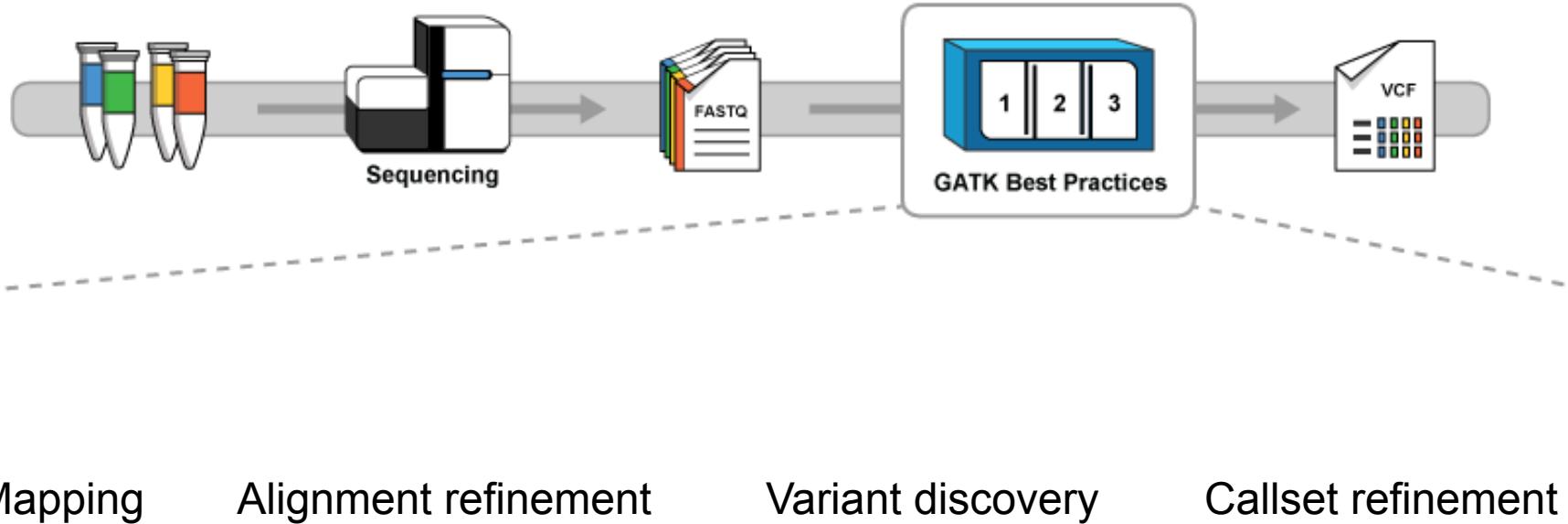
Removed sequence

Basic quality control - FASTQC

Module load FastQC



Genome Analysis Tool Kit (GATK) SciLifeLab

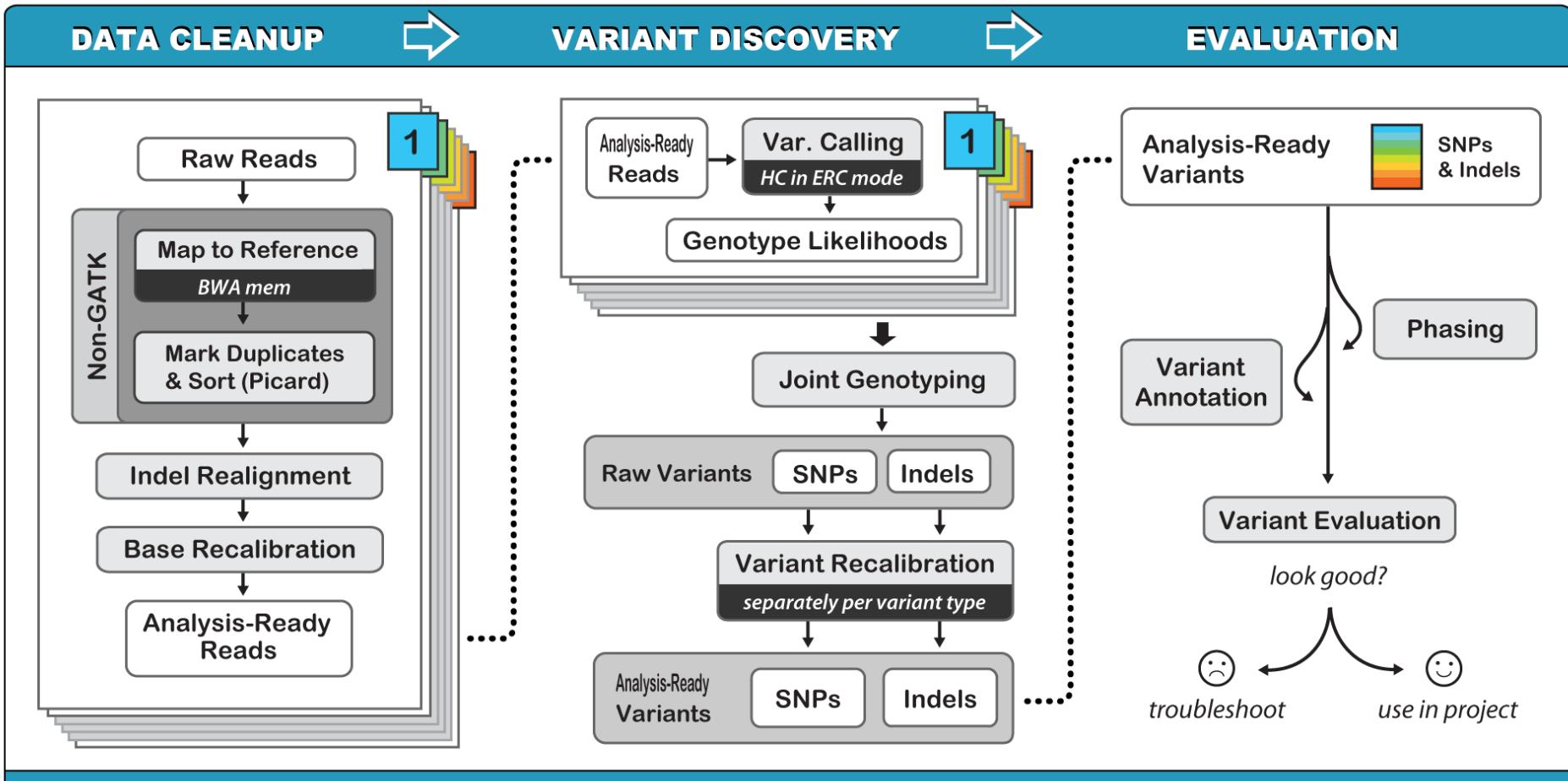


Mapping

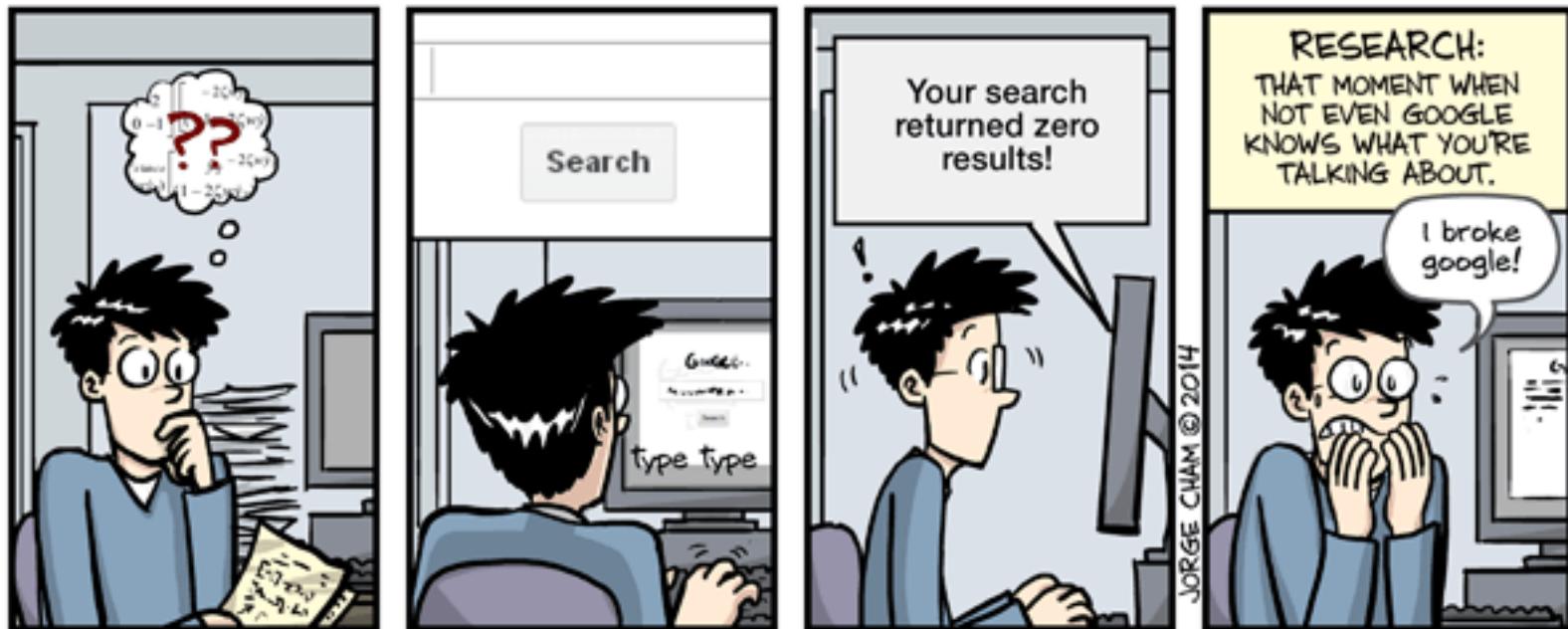
Alignment refinement

Variant discovery

Callset refinement

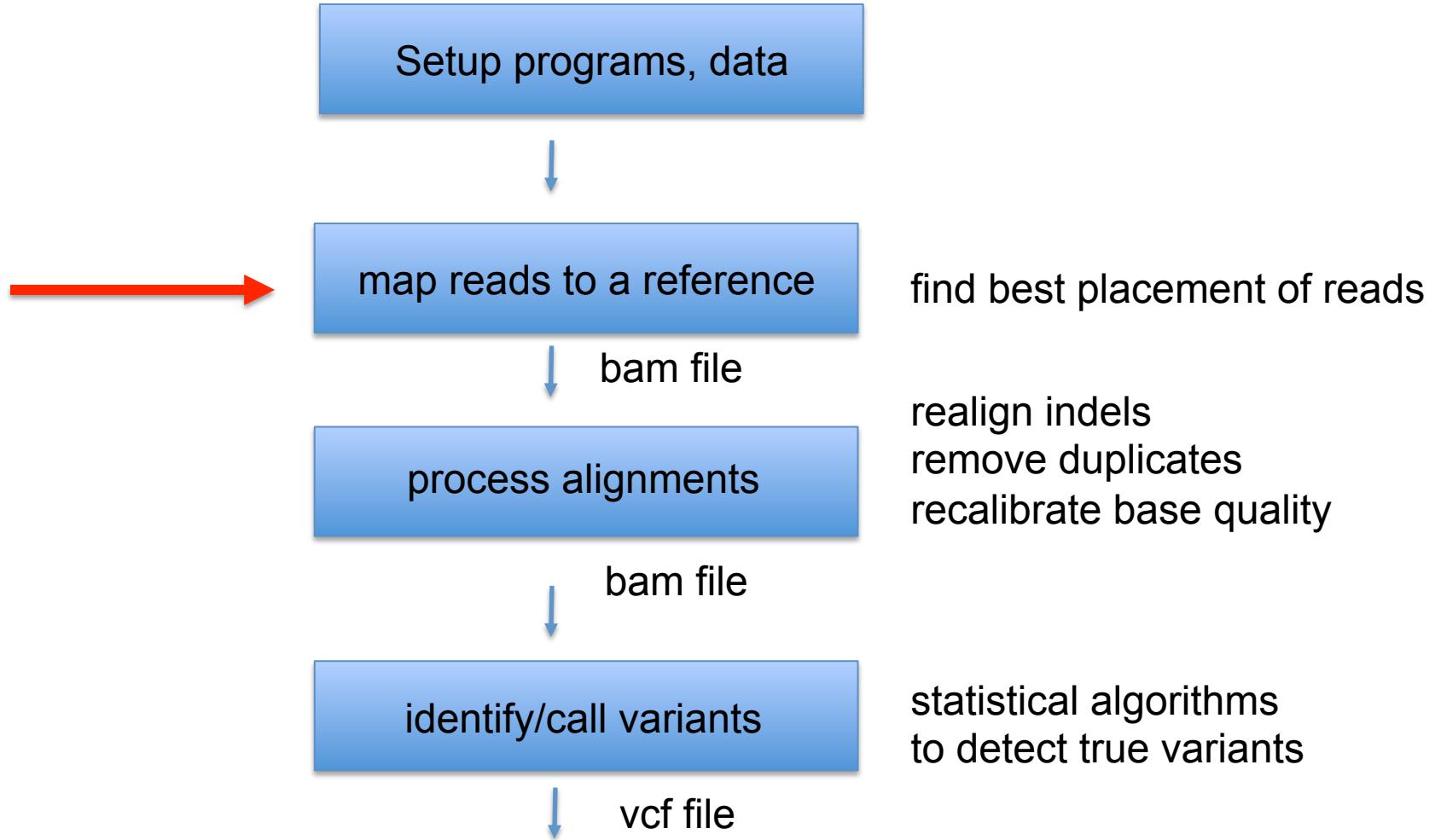


When in doubt, google it!



Steps in resequencing analysis

SciLifeLab



Mapping to reference genome

brute force

TCGATCC

x

GACCTCA**TCGATCC**CACTG

brute force

TCGATCC

x

GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
x
GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
x
GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
| | x
GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
x
GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
X
GACCTCA**TCGATCC**CACTG

brute force

TCGATCC
| | | | |
GACCTCA**TCGATCC**CACTG

hash tables

build an index of the reference sequence for fast access

	0	5	10	15	
seed length 7	GACCTCATCGATCCCACTG				
	GACCTCA	→	chromosome 1,	pos 0	
	ACCTCAT	→	chromosome 1,	pos 1	
	CCTCATIC	→	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
CCTCATIC	→	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
CCTCATIC	→	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

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 NA ^BANA ANANA ^B BANANA ^	ANANA ^B ANA ^BAN A ^BANAN BANANA ^ NANA ^BA NA ^BANA ^BANANA ^BANANA	ANANA ^B ANA ^BAN A ^BANAN BANANA ^ NANA ^BA NA ^BANA ^BANANA ^BANANA	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

Reference genome

Reference.fasta

Reference.fai

```
>Potra000002
CACGAGGTTCATCATGGACTTGGCACCATAAAA
GTTCTCTTCATTATATTCCCTTAGGTAAAATG
ATTCTCGTTCATTGATAATTGTAAATAACCGG
CCTCATTCAACCCATGATCCGACTTGATGGTGA
TACTTGTGTAATAACTGATAATTACTGTGATTT
ATATAACTATCTCATAATGGTCGTCAAAATCTT
TTAAAAGATAAAAAAACCTTATCAATTATCTA
TATAAAATTCAAATTGTACACATTTACTAGAAAT
TACAACTCAGCAATAAAATTGACAAAATATAAAA
CAGAACCGTTAAATAAGCTATTATTCATC
ACAAAACATCTAAGTCAAAATTGACATAAGTT
TCATCAATTACAAACAAACACAATTTCACAAAAA
TCTCAACCAACCATAACATGTACAAATTATAAA
TATCAACAATTGTTGAGAAAAAACTATAAC
ACAAGTAAATACCAAAAAAAATACATATACTACA
AAACAATATAAAAAATTACATTTAAAATTG
TGTCAAATAAAAATTAGATTGCTTACTTAAG
CTGGAGAATTGCAATAAAATTGCAATTAGAAGA
```

Sample data

R1.fastq

R2.fastq

```
@HISEQ:100:C3MG8ACXX:
5:1101:1160:2197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFFHHHHGJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIIJJJJIIJJJJIIIEIHHIJ
HGHHHHDFFFEDDDDCDDDCDDDDDCDC
@HISEQ:100:C3MG8ACXX:
5:1101:1448:2164 1:N:0:ATCACG
NAGATTGTTGTGCCTAAATAAAATAAAATAAAAT
AAAAATGATGATGGCTTAAAGGAATTGAAATT
AAGATTGAGATATTGAAAAAGCAGATGTGGTC
+
#1=DDFFEHDFHHJGGIJJJGJIHIGIJJJJI
IJJJJIJJFJJF?
FHHIIJJJJGJIJJJIJIGHGHIIJJJIHGH
GUCHEEEEDEEE>GDDD
```

Output from mapping



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.ILLUMINA.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	
AGTAAAGACCCCTGCCCAAGCAAAGGATA			AAAGGAAATGA	AGTTTGAATAATA	?@?;@@ABAB8@<?B@B;A@@@B@@A>A@>>:<8A@@B@@@B@AA@@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	
AACATAACACATCTCTATCAGAATTAGAGTA			AAAGACCCCTGCCCAAGCAAAGGAT		>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	
AATGATAGAAAAGCACAAATTGTAACGAAAGAACGCTCGAAATC			TCGCATCCTCCTGAC		@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

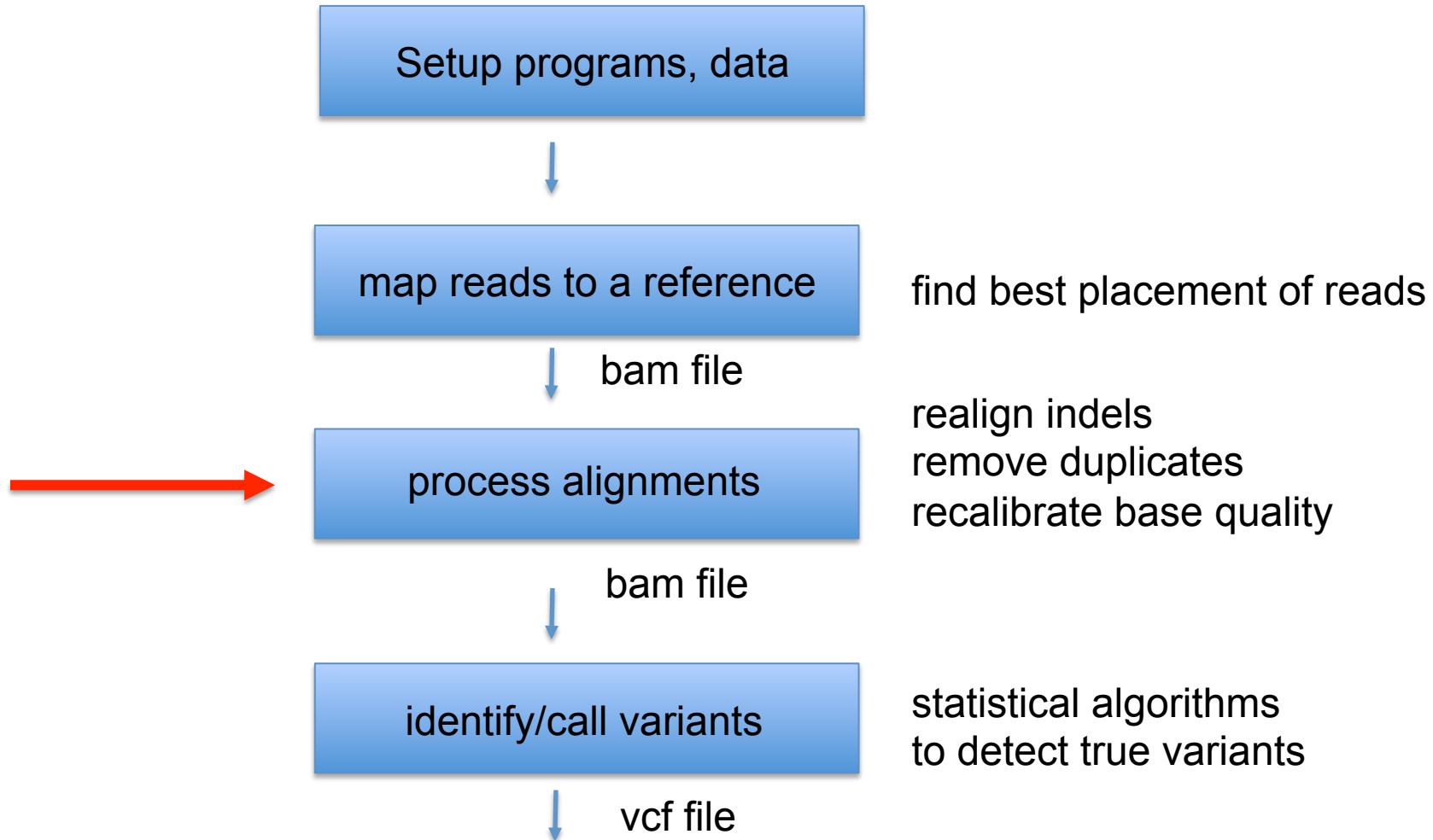
Chr

Start position

Sequence

Quality

Steps in resequencing analysis



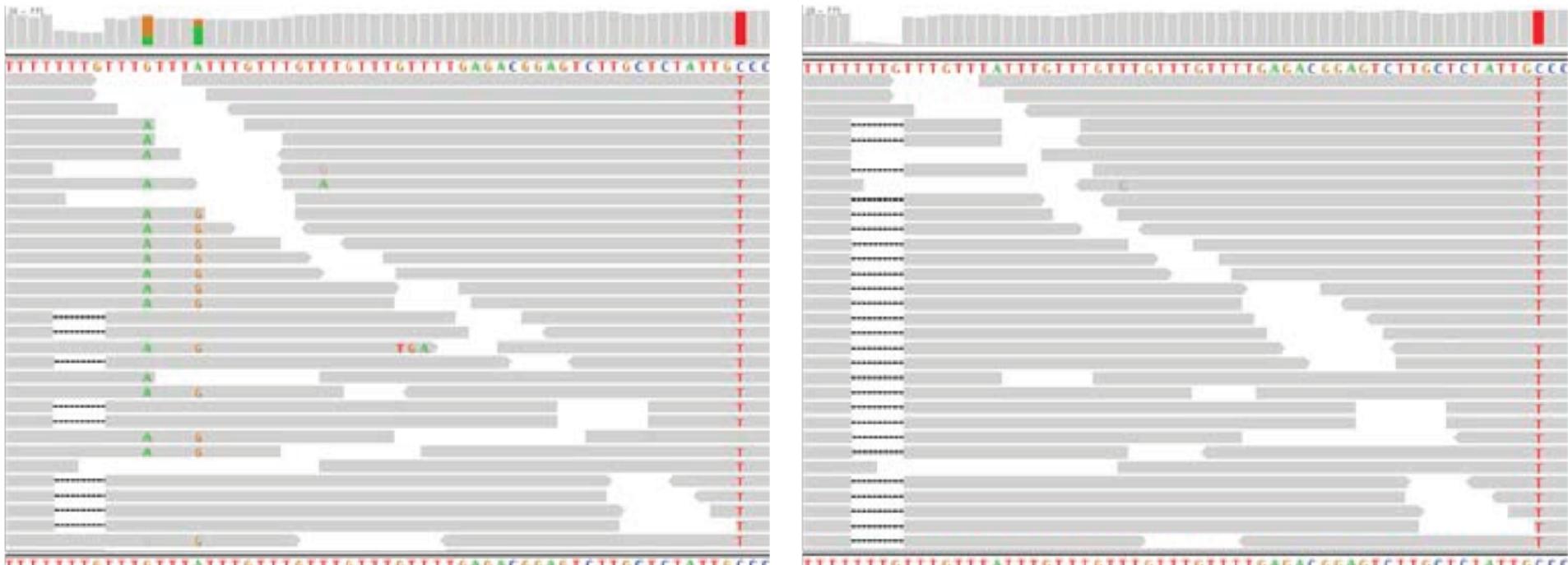
Processing BAM files

.bam

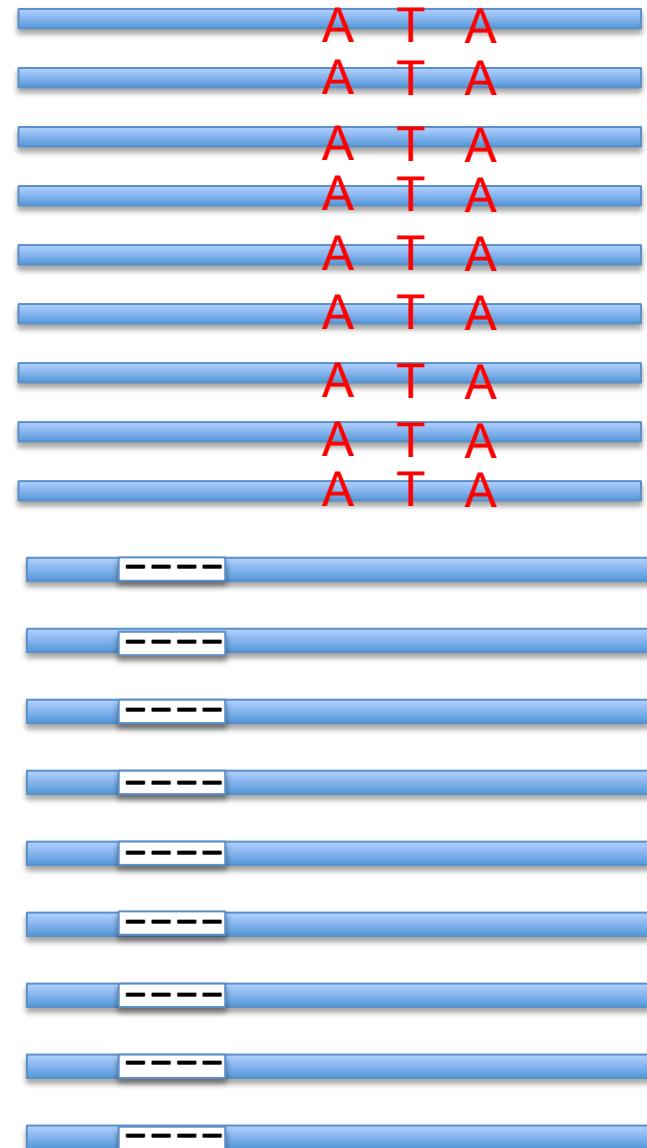
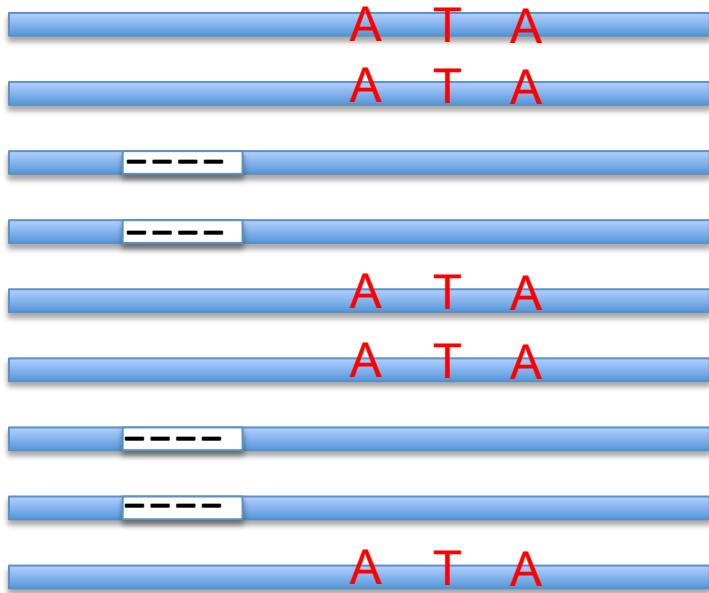


Realign around indels

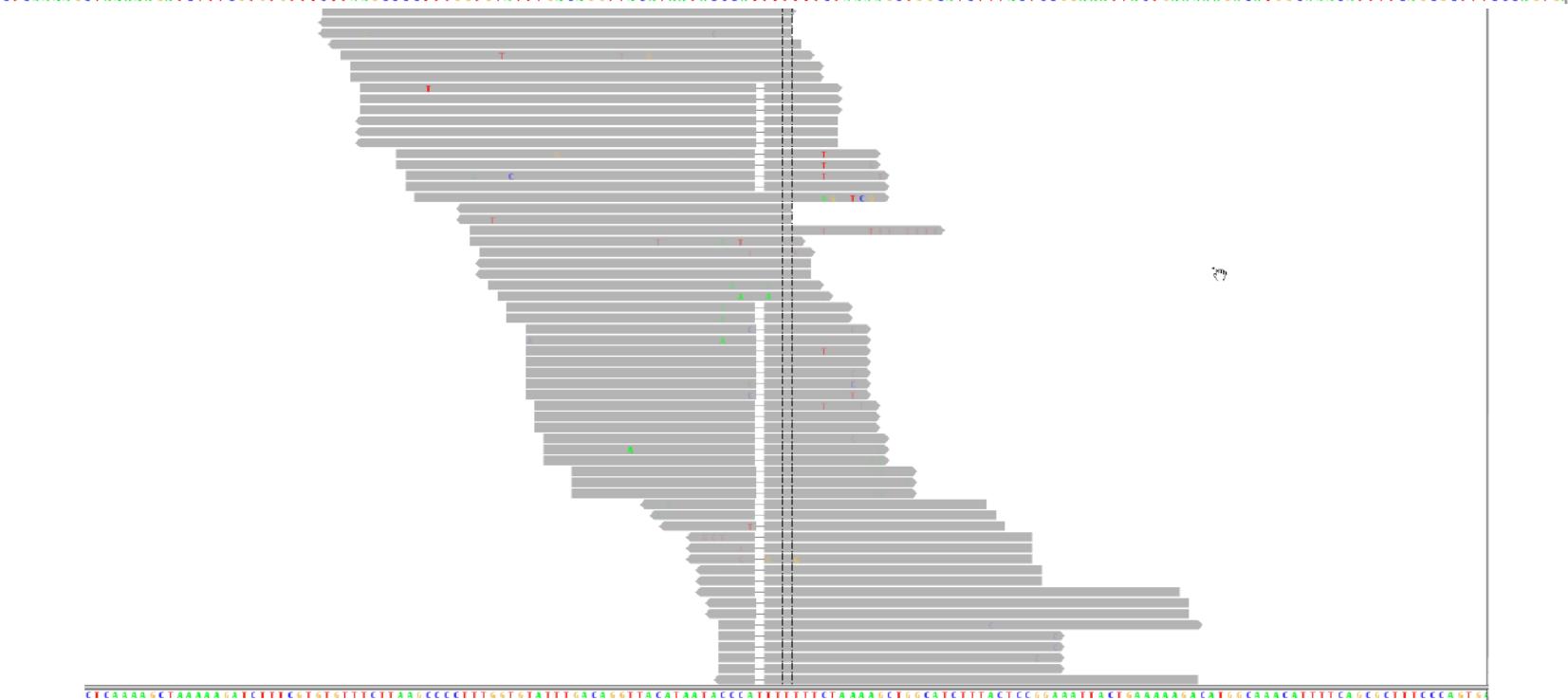
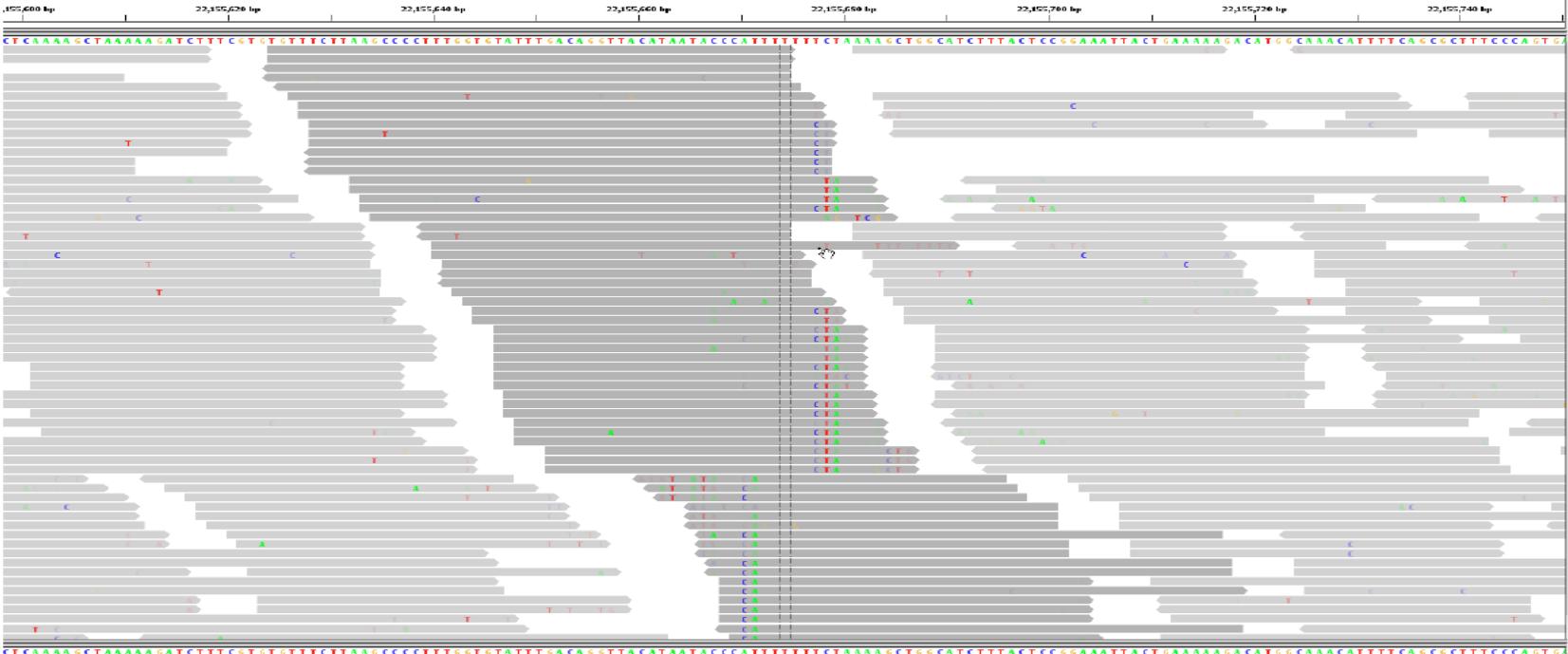
- mapping is done one read at a time
- single variants may be split into multiple variants
- solution: realign these regions taking all reads into account



Local realignment

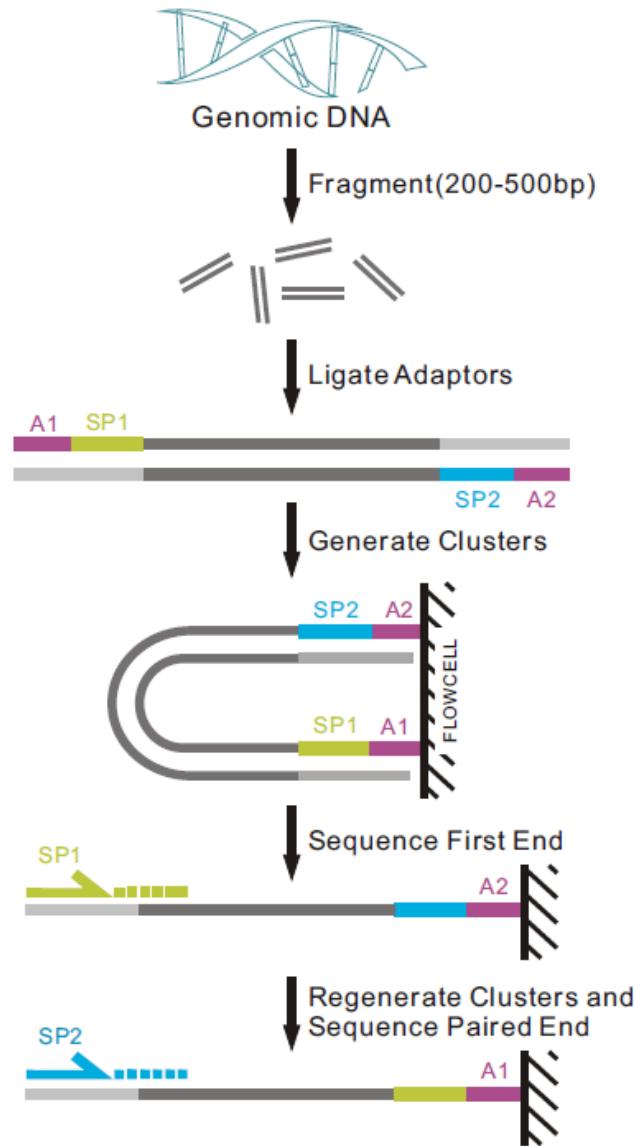


can be performed using GATK commands:
RealignerTargetCreator followed by
IndelRealigner



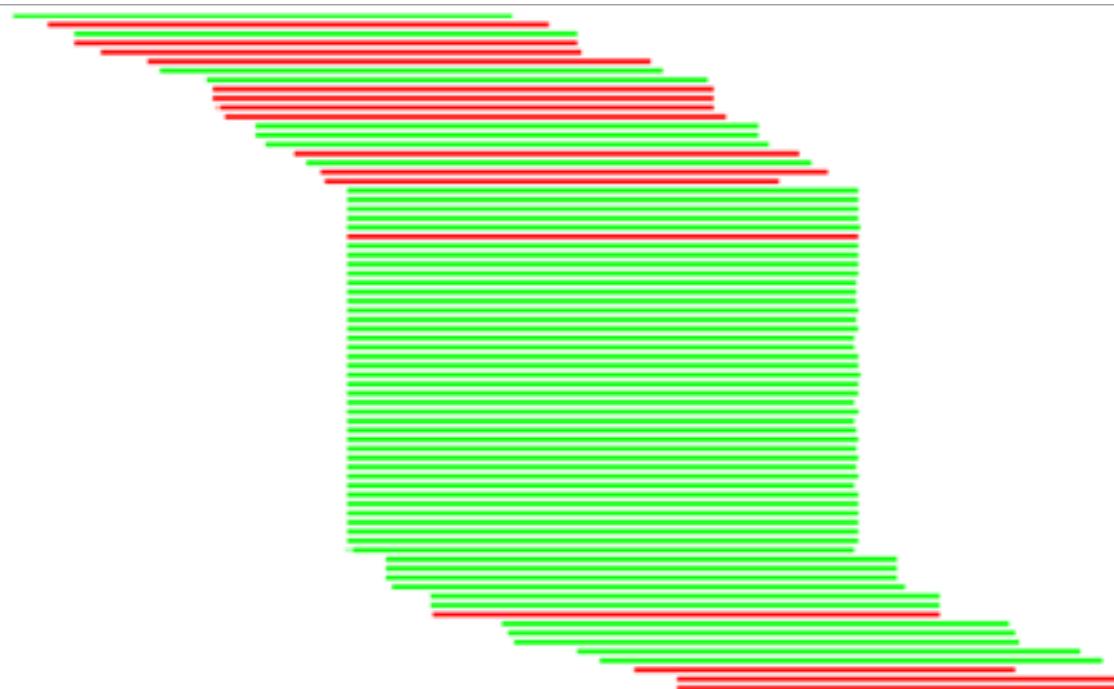
- When two or more reads originate from same molecule (artificial duplicates)
 - not independent observations
 - skew allele frequency and read depth
 - errors double counted
- PCR duplicates occur
 - during library prep, or
 - optical duplicates (one cluster read as two)
- mark or remove
- Reading: <http://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/>

Paired end sequencing



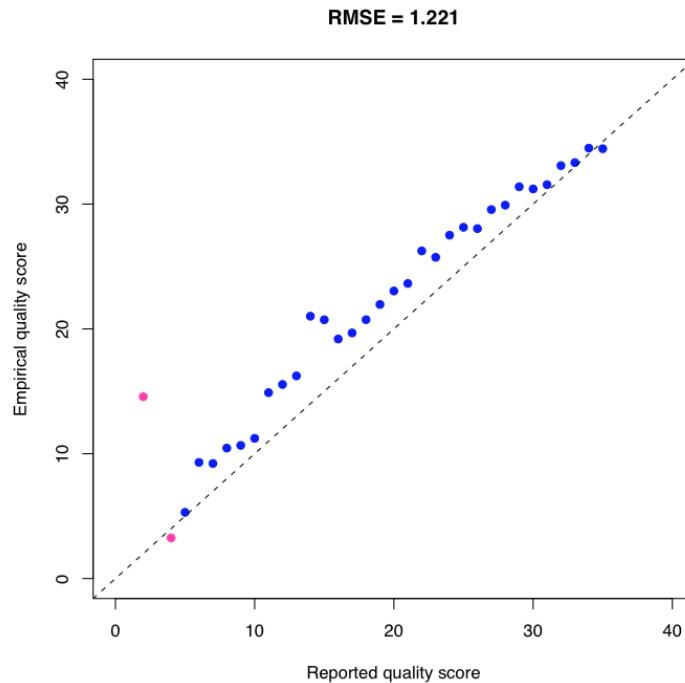
Identify PCR duplicates

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

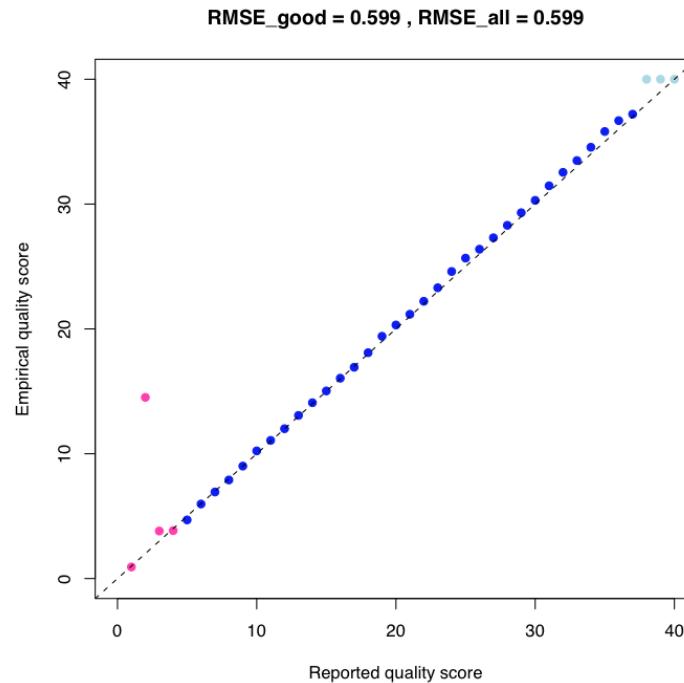


Base quality recalibration

Reported Quality vs. Empirical Quality



Original Data



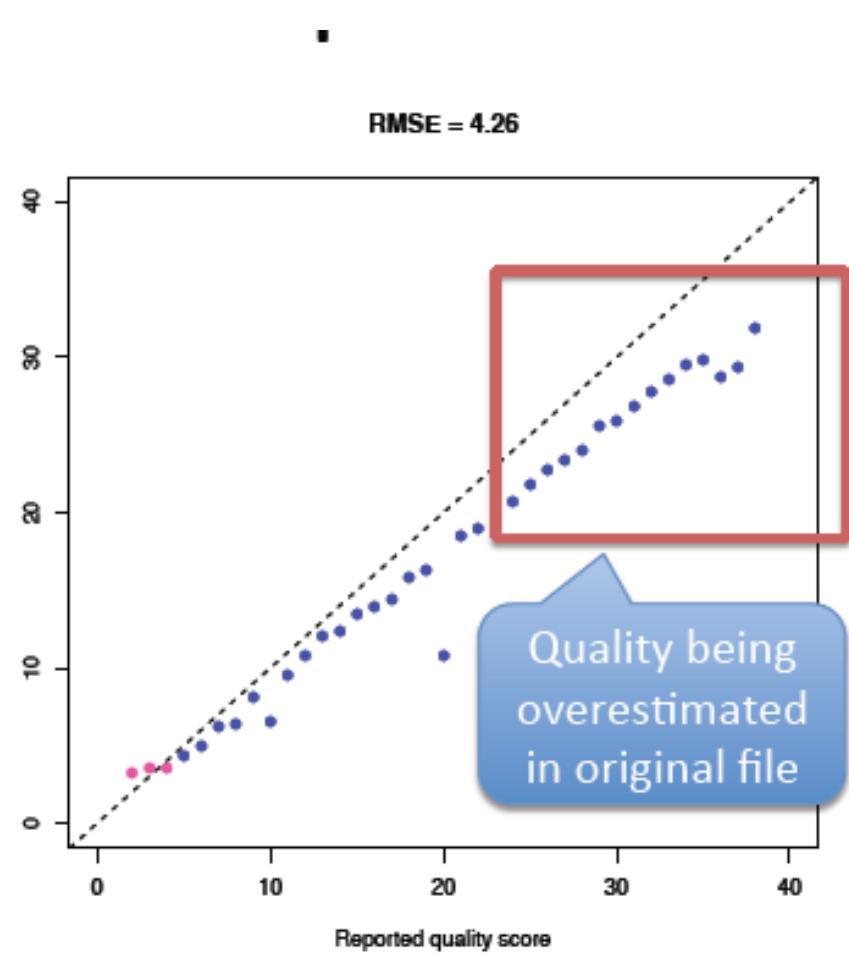
After GATK Recalibration

Recalibration Method

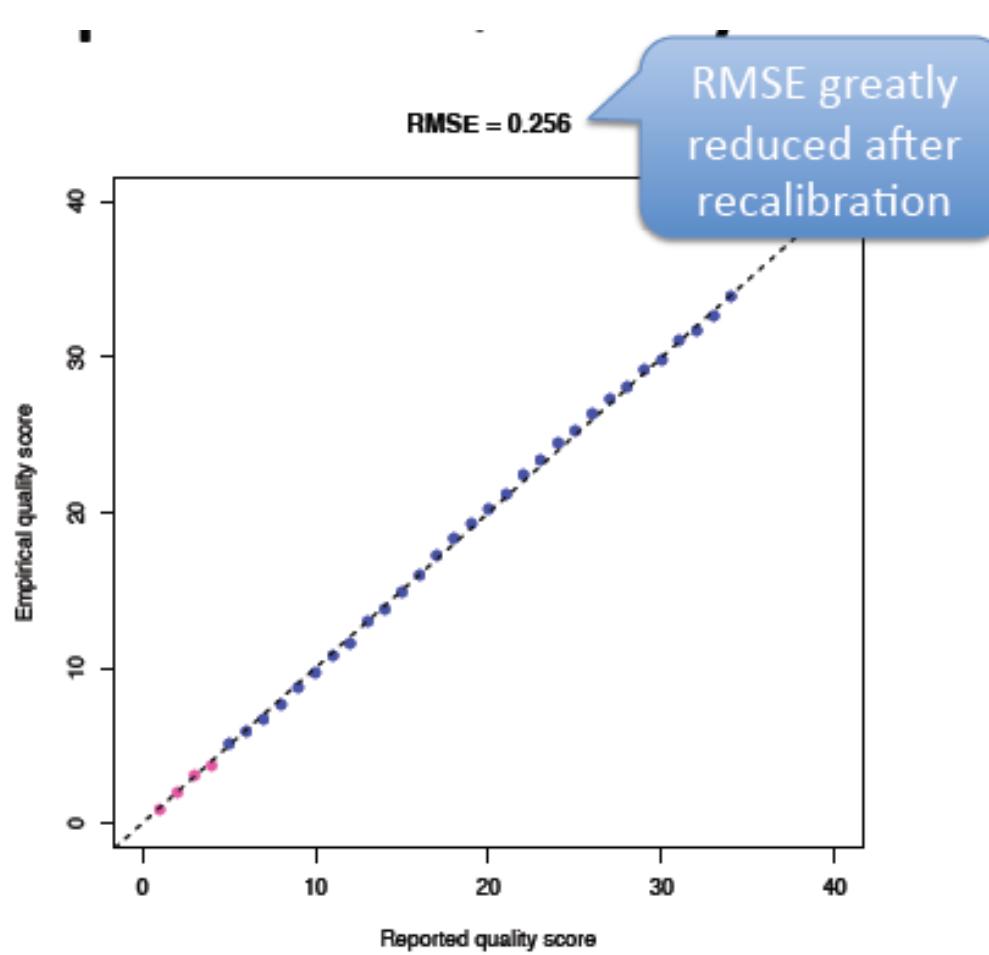
- Bin each base by
 - read group
 - called quality
 - position in read
 - local dinucleotide context
- score observed quality per bin
 - # of mismatches +1 / # of observed bases
- scale compared to reported quality
- Further reading:
- <http://gatkforums.broadinstitute.org/gatk/discussion/44/base-quality-score-recalibration-bqsr>

Reported vs empirical quality scores

SciLifeLab



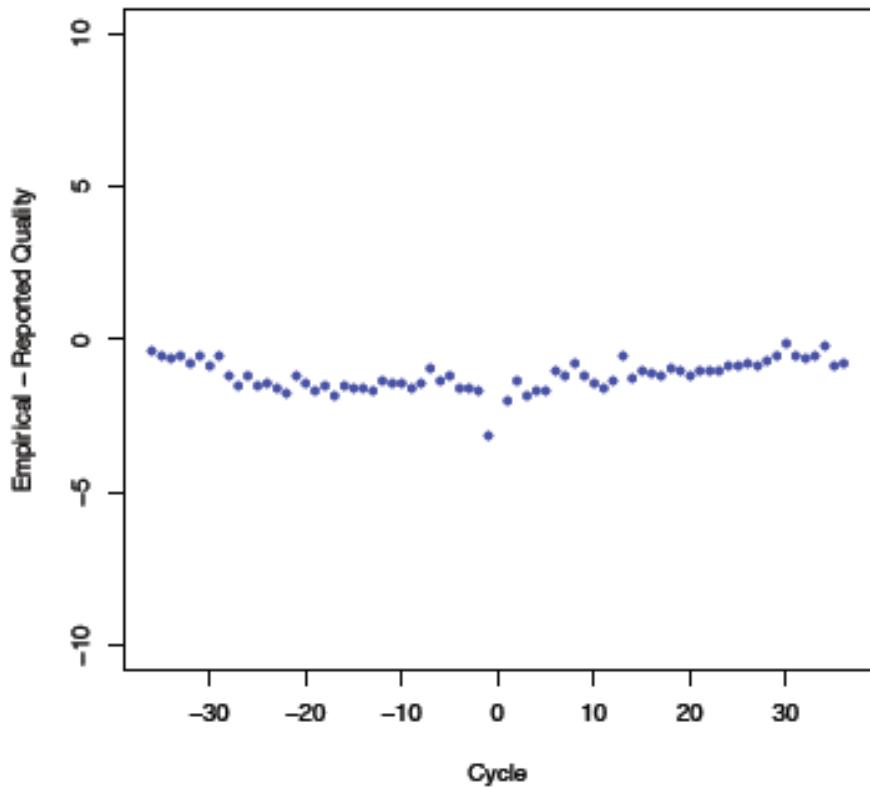
Before Recalibration



After Recalibration

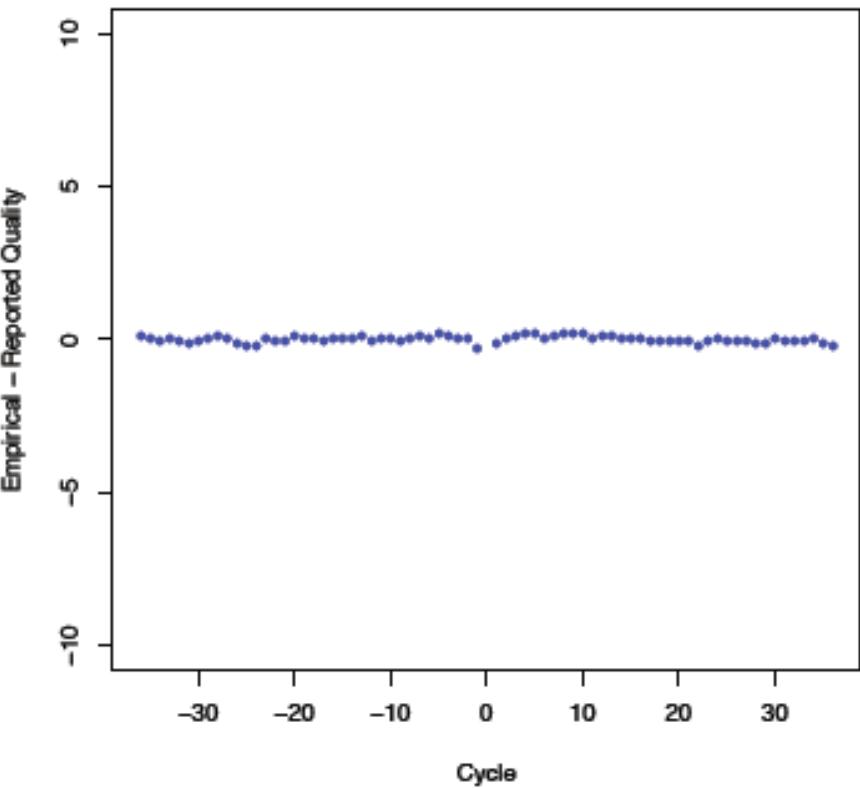
Residual error by machine cycle

RMSE = 1.275



Before Recalibration

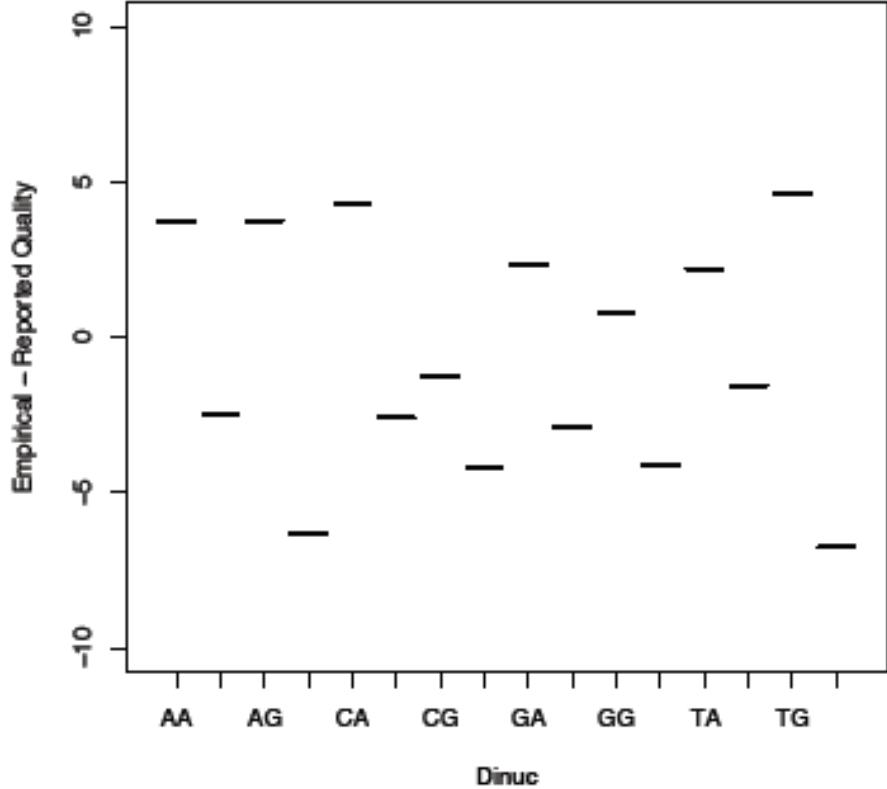
RMSE = 0.105



After Recalibration

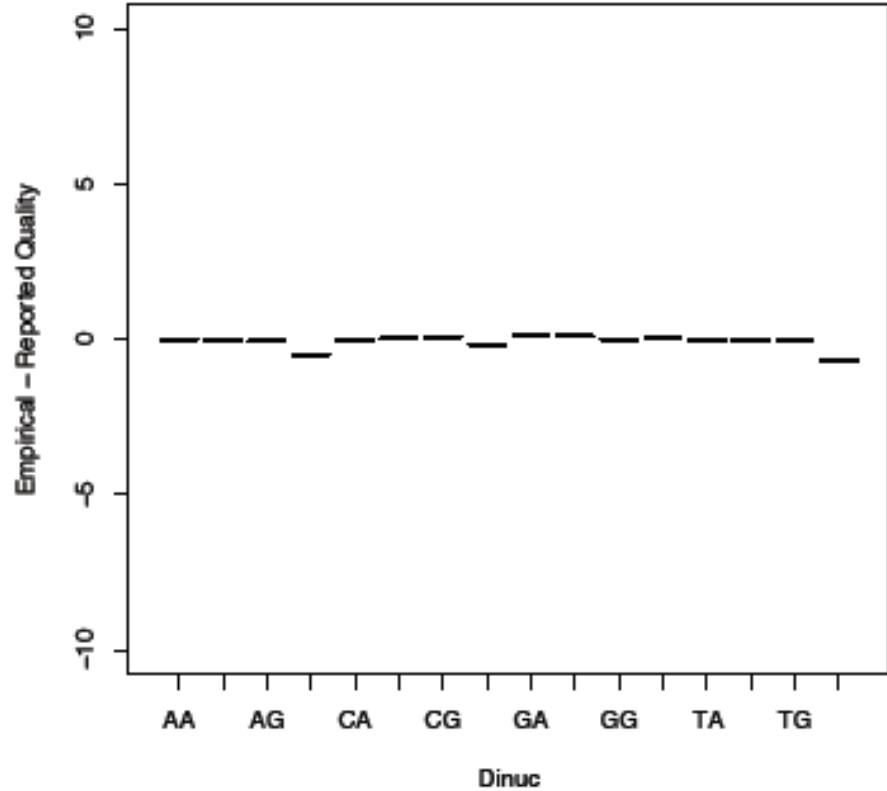
Residual error by dinucleotide

RMSE = 4.188



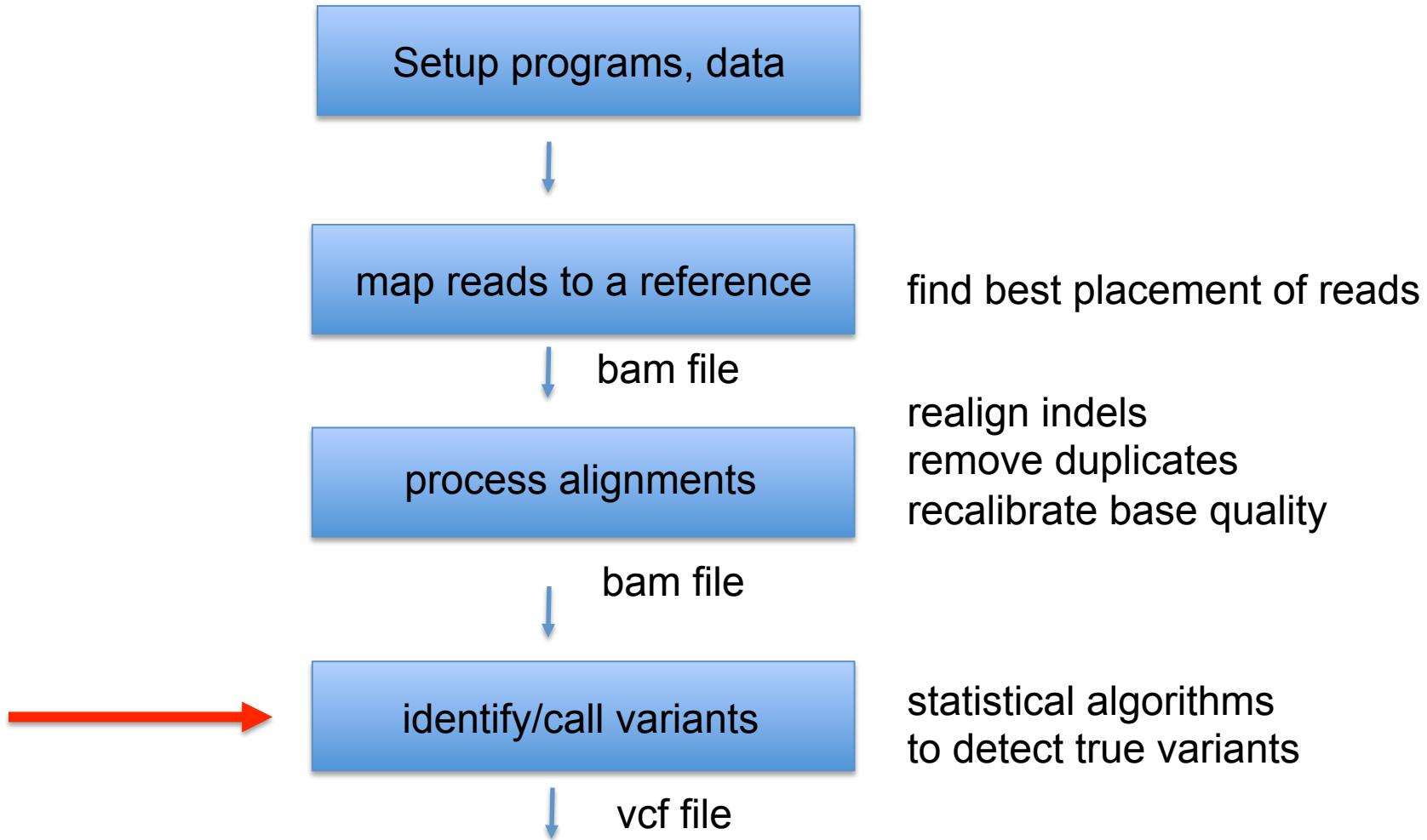
Before Recalibration

RMSE = 0.281



After Recalibration

Steps in resequencing analysis



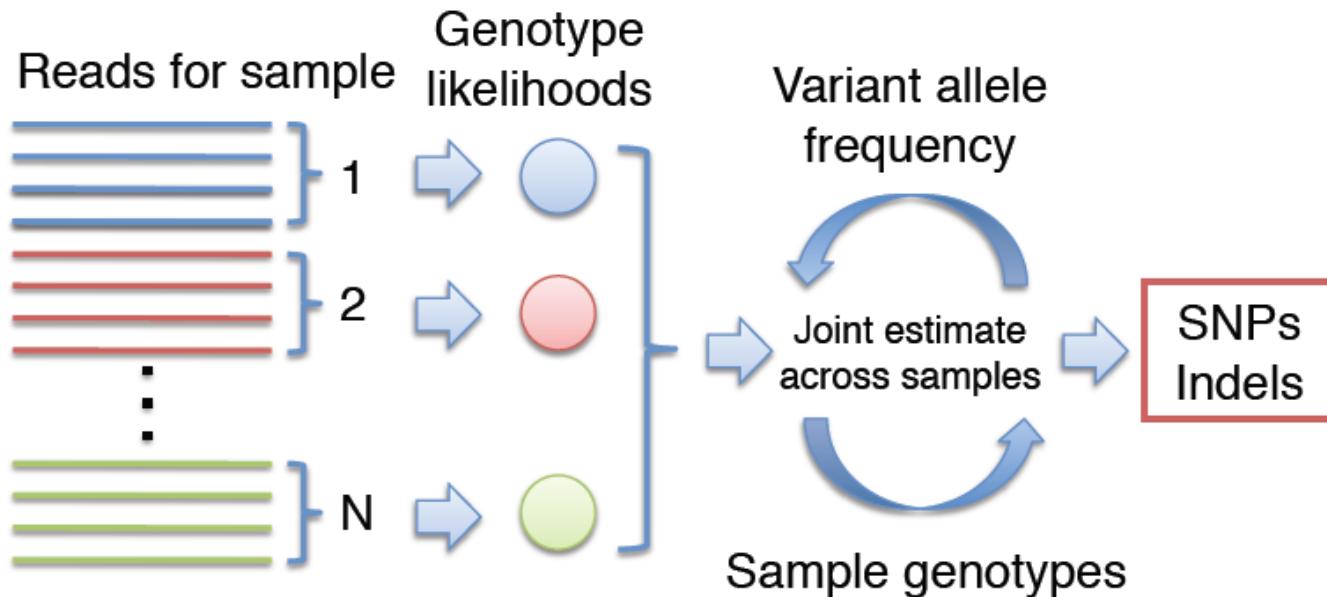
Variant calling

simple pileup methods

Reference: acacagatagacatagacatagacagatgag

acacagatagacatagacatagacagatgag
acacacatagacatagacatagacagatgag
acacagatagacatagacatagacagatgag
acacagatagacatatacatagacagatgag
acacagatagacatatacatagacagatgag
acacagatagacatatacatagacagtgag
acacagatagacatagacatagacagatgag
acacagatagacatatacatagacagatgag
acacagatagacatagacatagacagatgag

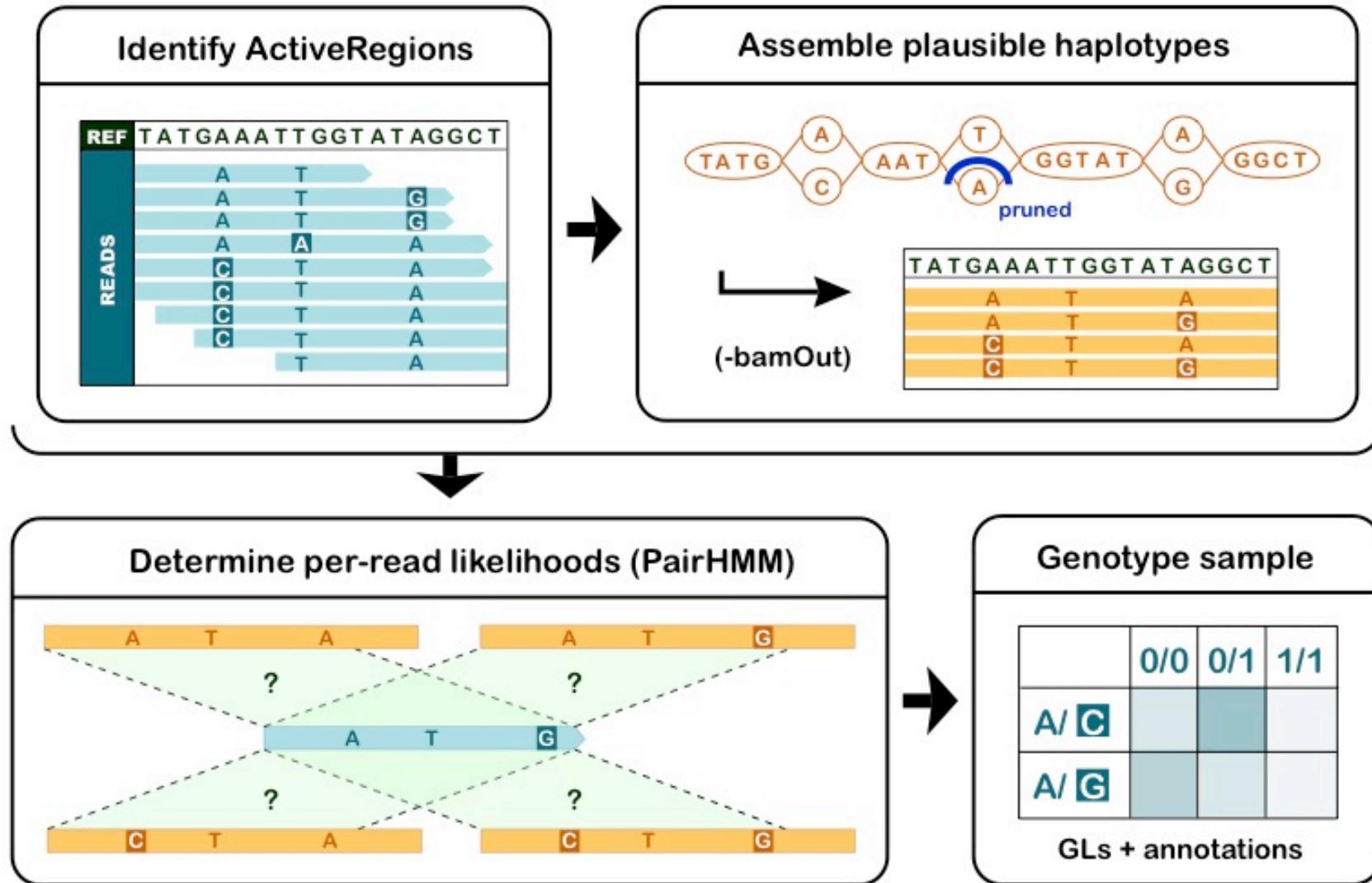
Baysian population-based calling



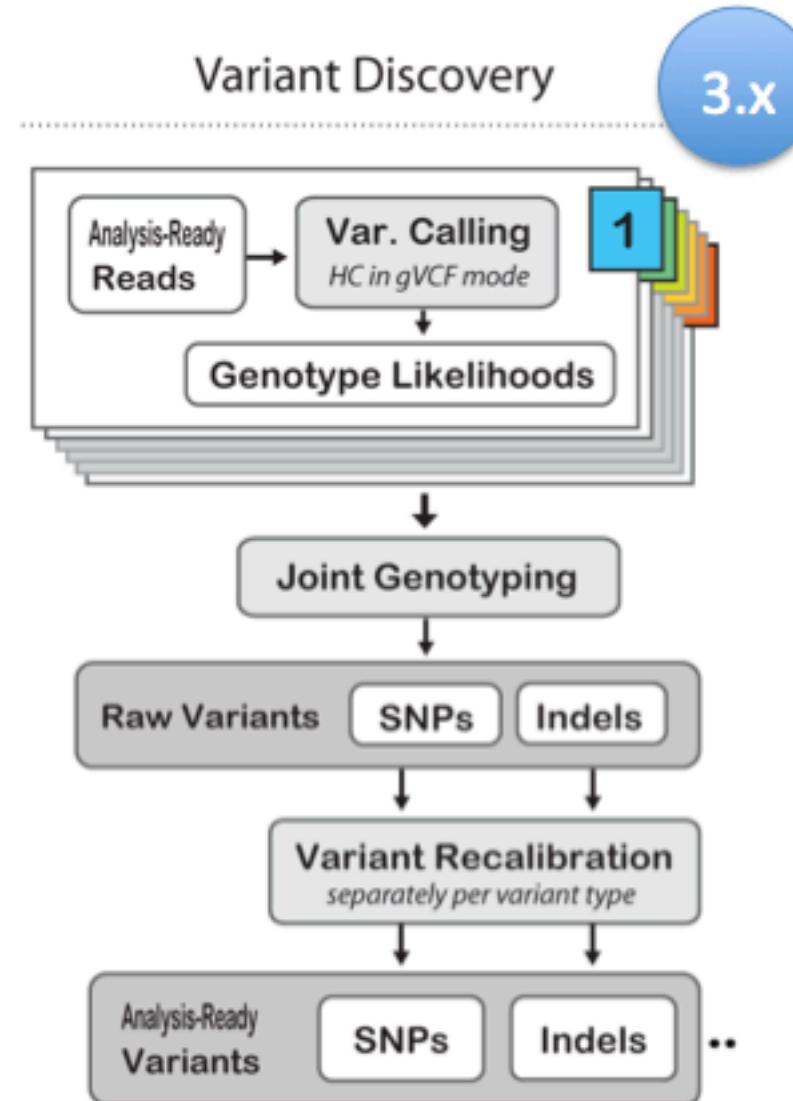
Simultaneous estimation of:

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

GATK haplotype caller



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
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

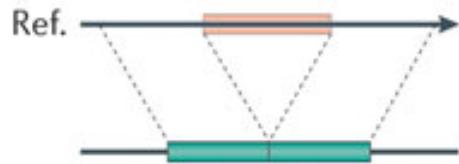
```
##fileformat=VCFv4.0 ##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth"> ##INFO=<ID=AF,Number=.,Type=Float,Description="Allele 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="Less than 10% quality samples">
##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
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|2:21:6:23,27
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=50,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 14070 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

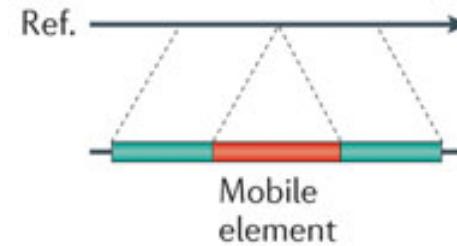
Deletion



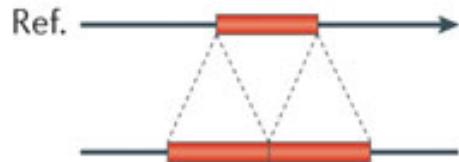
Novel sequence insertion



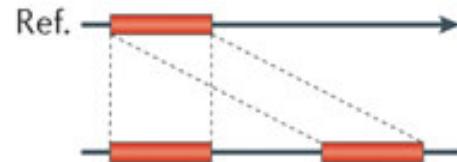
Mobile-element insertion



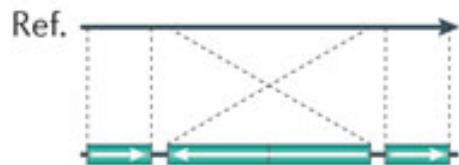
Tandem duplication



Interspersed duplication



Inversion



Translocation



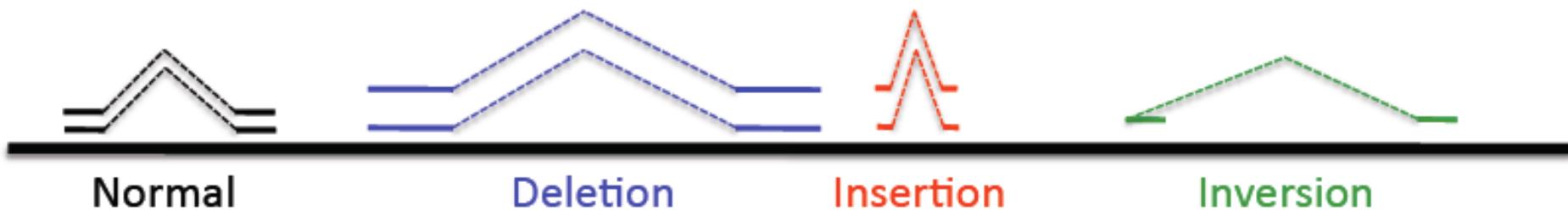
1) Read depth analysis

- 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.



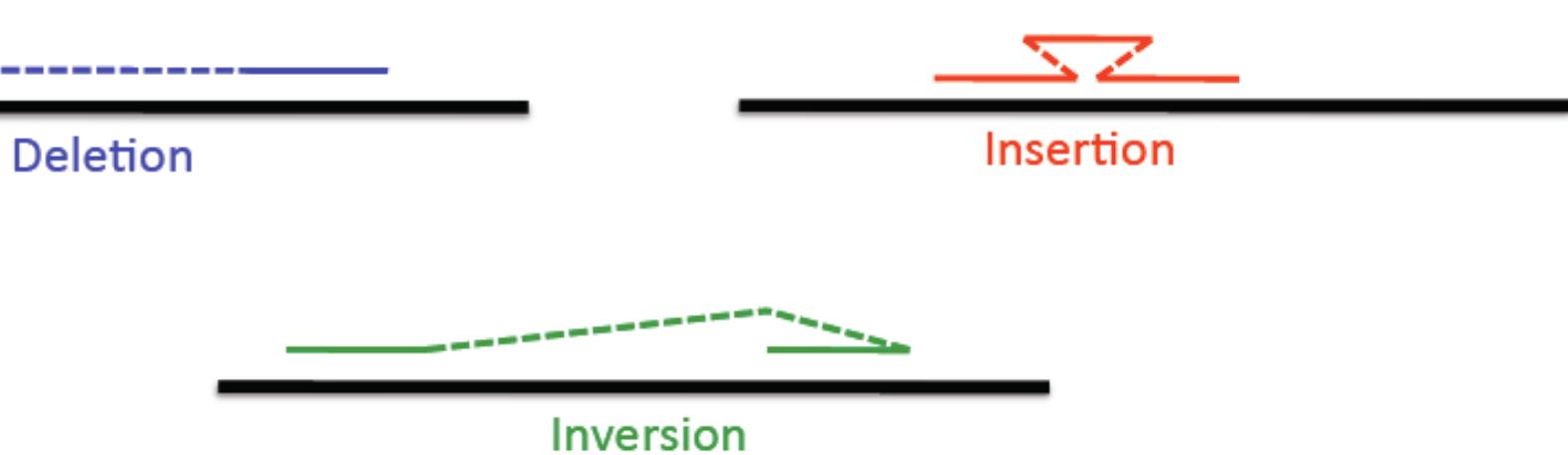
2) Paired end analysis

- 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



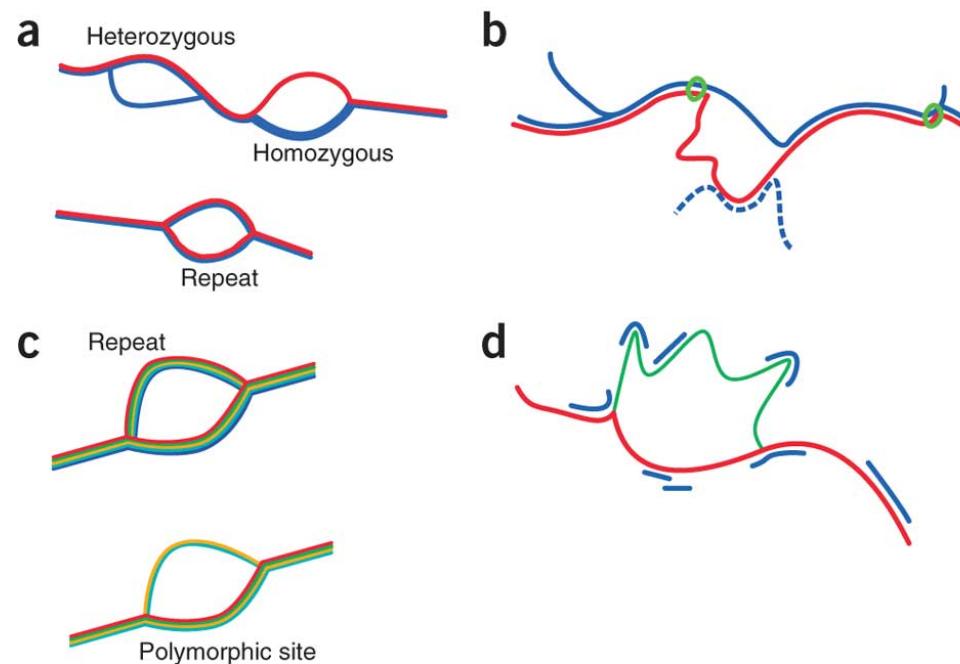
3) Split-read alignments

- 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



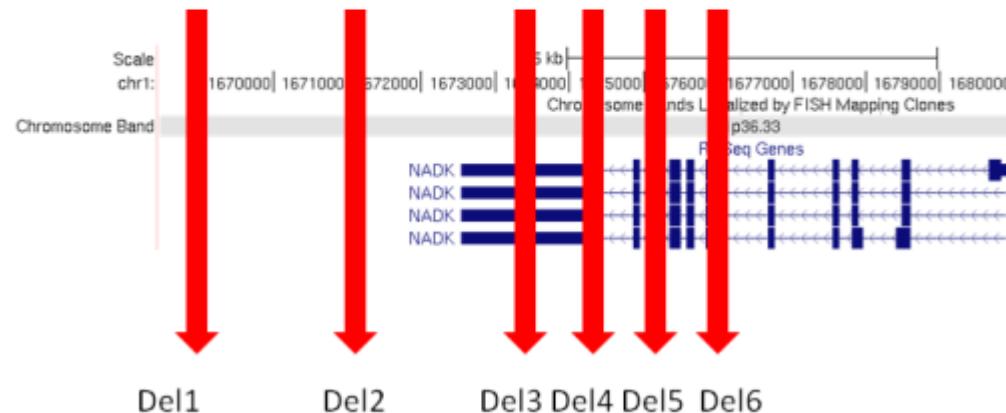
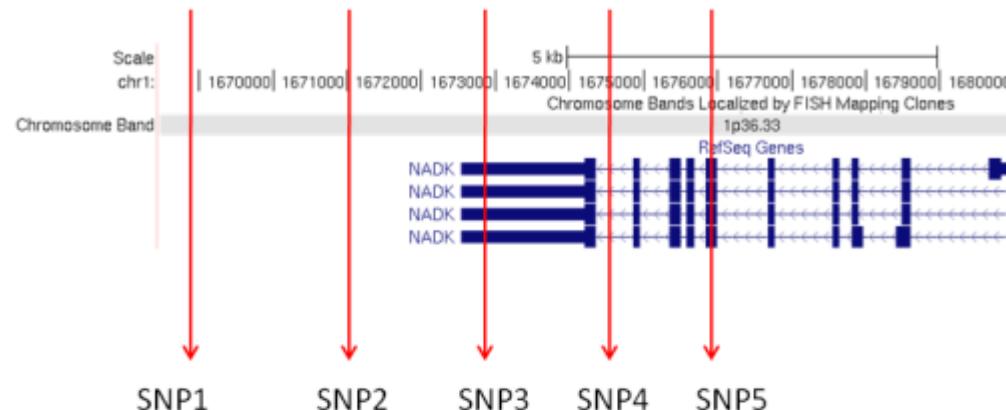
4) *De novo* assembly to identify structural variants

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



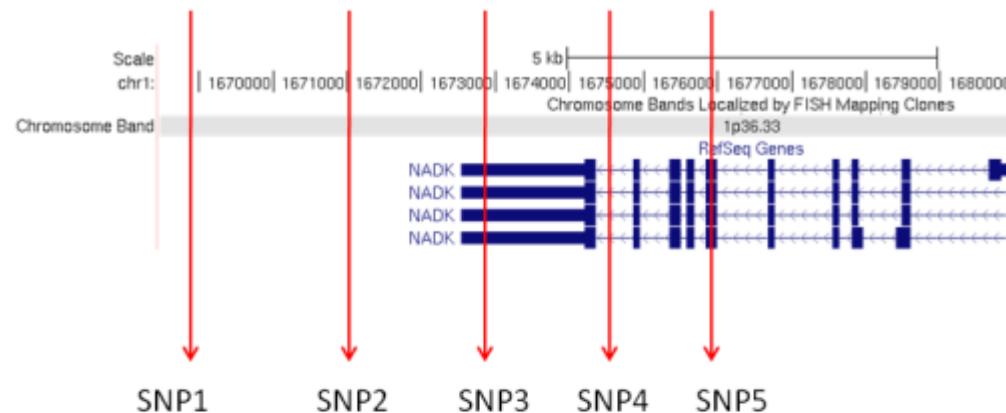
Annotation of variants

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



Annotation of variants

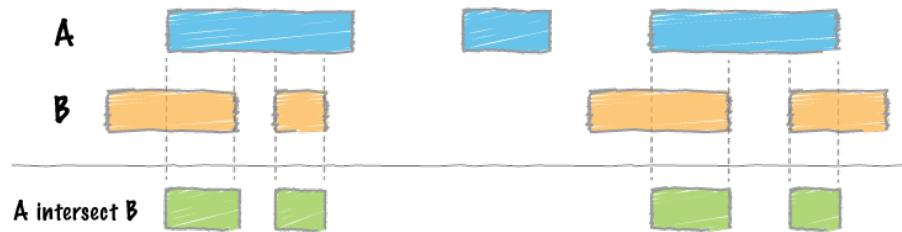
By comparing with existing annotation for the reference genome it is possible to gain information about localization and expected effect



Most commonly used tools are Annovar and SNPEff

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 exercise

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

- 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, recalibration, call and filter variants

1) Access to data and 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/
```

Sample1: NA06984

Sample2

read1.fq

read2.fq

read1.fq

read2.fq

Mapping**Process alignments****Genotyping****Joint genotyping****Filtering**bwa aln
bwa alnbwa aln
bwa aln

NA06984.realign.dedup.recal.bam

NA06984.realign.dedup.recal.bam

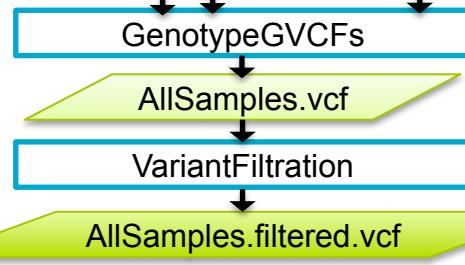
HaplotypeCaller

HaplotypeCaller

NA06984.g.vcf

Sample2.g.vcf

Sample3.g.vcf



View in IGV

2) Align each paired end separately

```
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

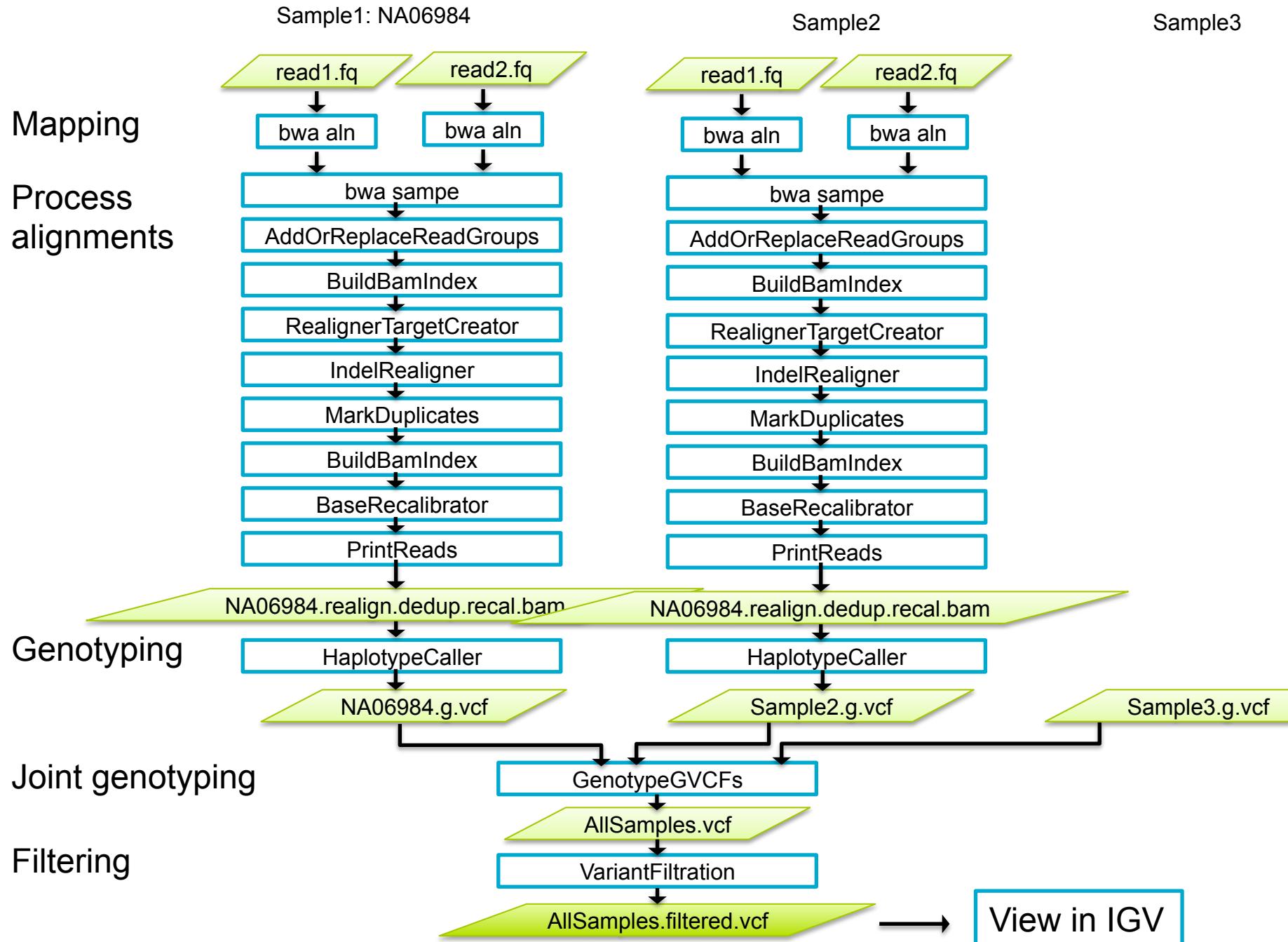
```
java -Xmx2g -jar /path/GenomeAnalysisTK.jar  
-T VariantFiltration  
-R <reference>  
-V <input vcf>  
-O <output vcf>  
--filterExpression "QD<2.0" --filterName QDfilter  
--filterExpression "MQ<40.0" --filterName MQfilter  
--filterExpression "FS>60.0" --filterName FSfilter  
--filterExpression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

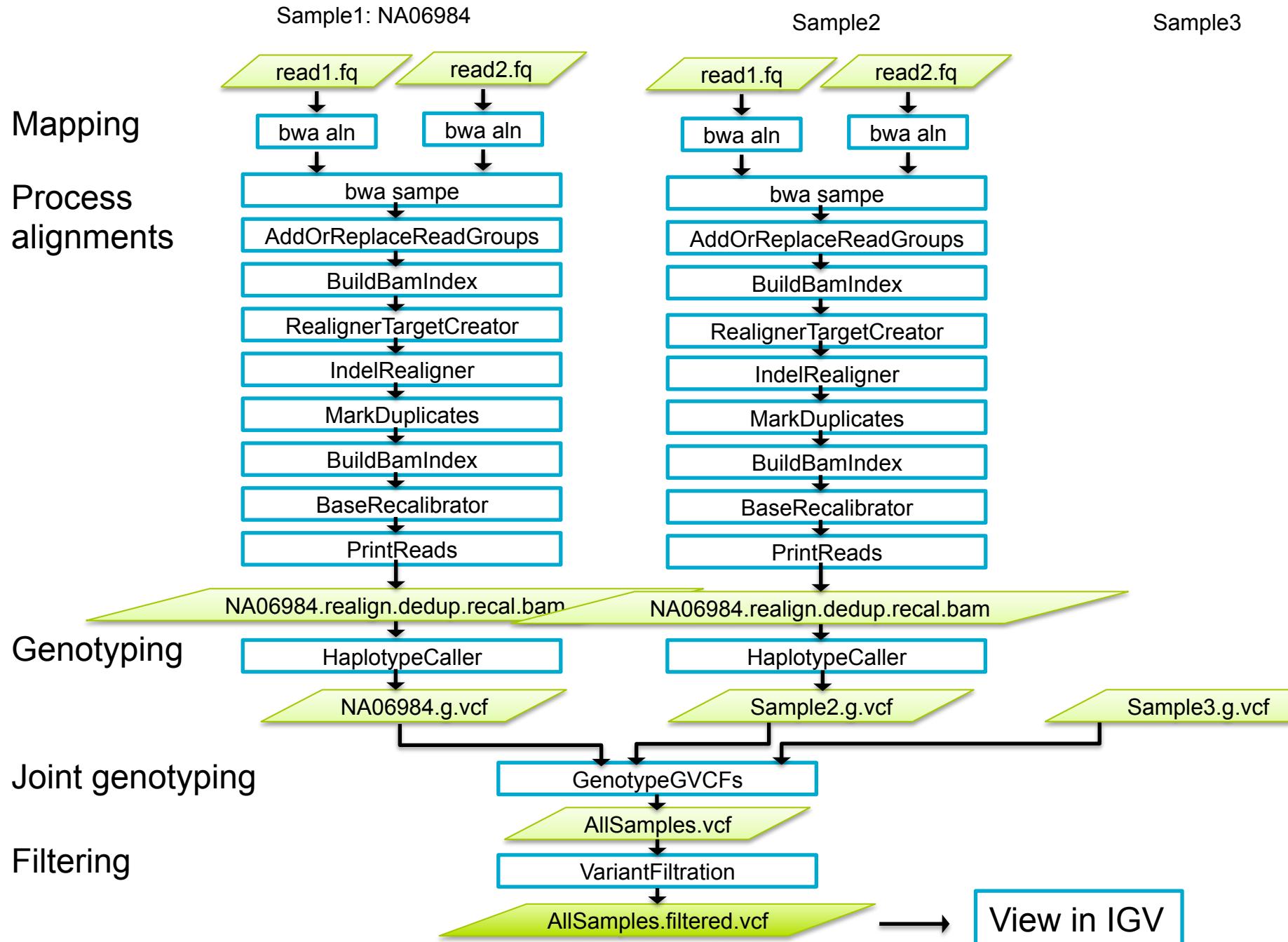
7) Viewing data with IGV

SciLifeLab

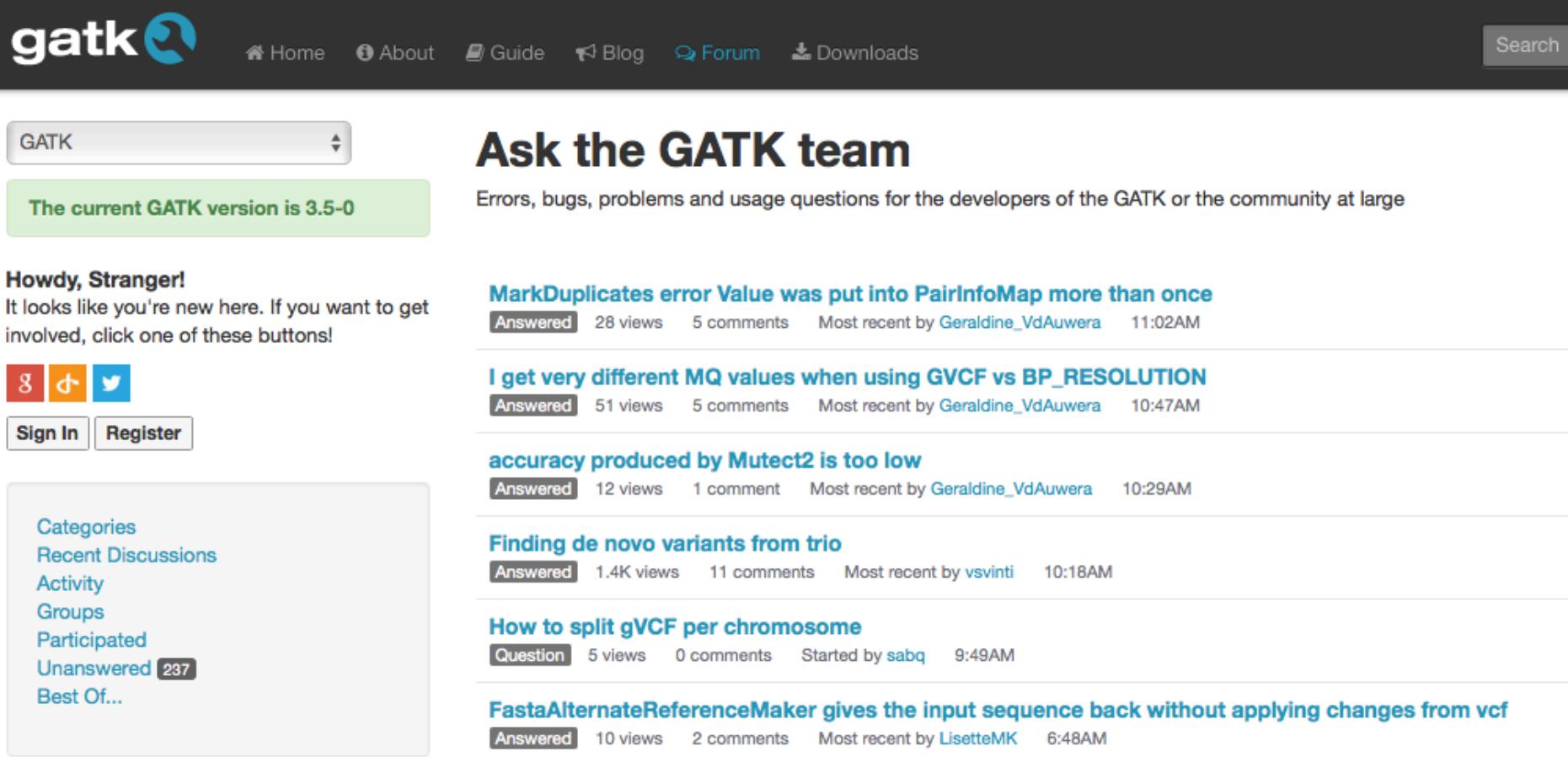


<http://www.broadinstitute.org/igv/>





- <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 screenshot shows the GATK Support Forum homepage. At the top, there's a navigation bar with links for Home, About, Guide, Blog, Forum (which is highlighted in blue), and Downloads. A search bar is also present. Below the navigation, a banner indicates the current GATK version is 3.5.0. On the left, a sidebar greets new users and provides links for Sign In and Register, along with social sharing icons for GitHub, Stack Overflow, and Twitter. The main content area features a section titled "Ask the GATK team" where users can ask errors, bugs, problems, and usage questions. Several support topics are listed, each with a title, status (Answered or Question), views, comments, and the most recent responder and timestamp. The topics include:

- MarkDuplicates error Value was put into PairInfoMap more than once** (Answered, 28 views, 5 comments, Most recent by Geraldine_VdAuwera at 11:02AM)
- I get very different MQ values when using GVCF vs BP_RESOLUTION** (Answered, 51 views, 5 comments, Most recent by Geraldine_VdAuwera at 10:47AM)
- accuracy produced by Mutect2 is too low** (Answered, 12 views, 1 comment, Most recent by Geraldine_VdAuwera at 10:29AM)
- Finding de novo variants from trio** (Answered, 1.4K views, 11 comments, Most recent by vsvinti at 10:18AM)
- How to split gVCF per chromosome** (Question, 5 views, 0 comments, Started by sabq at 9:49AM)
- FastaAlternateReferenceMaker gives the input sequence back without applying changes from vcf** (Answered, 10 views, 2 comments, Most recent by LisetteMK at 6:48AM)

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

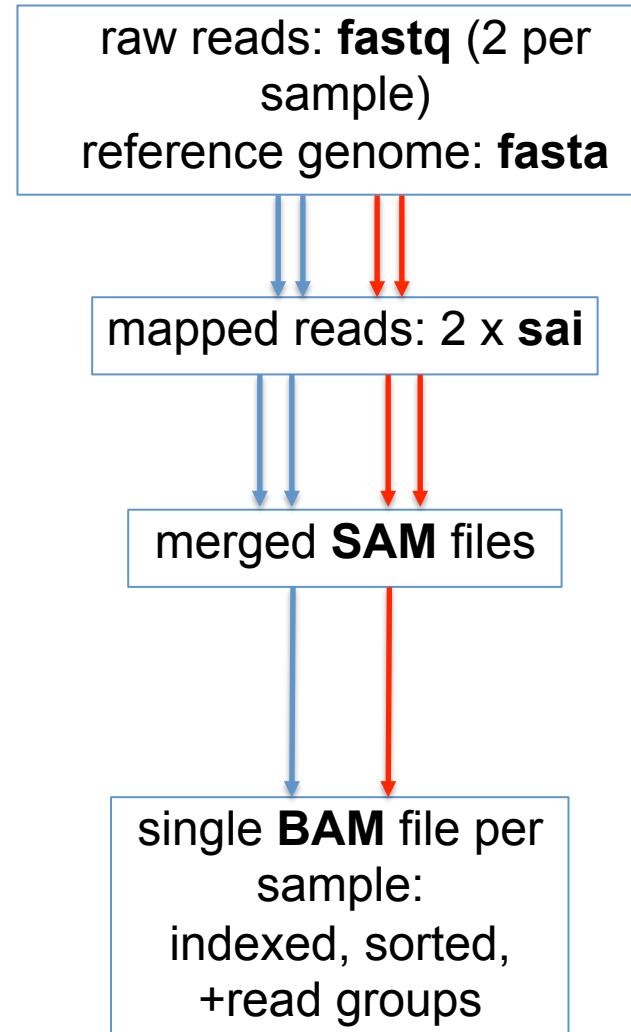
- `bwa index`
- `samtools faidx`
- `bwa aln`

3. Merging alignments

- `bwa sampe`

4. Creating BAM files

- `picard AddOrReplaceReadGroups`
- `picard BuildBamIndex`



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

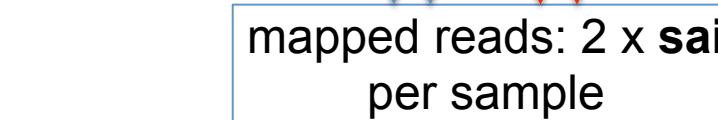
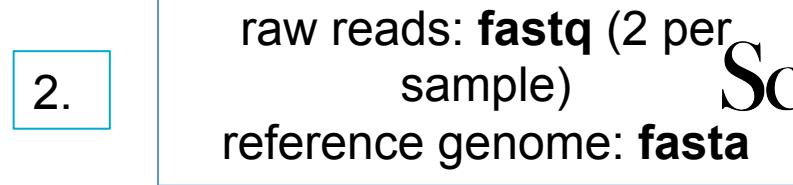
merged **BAM** file:
+realigned around indels
+mark/remove duplicates
+quality recalibrations

VCF file:
+filtered variants

mapping

processing

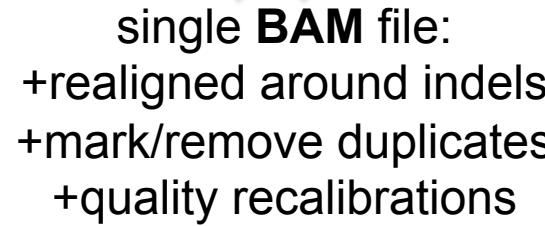
variant calling



4.



5.



6.



Naming conventions

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

...