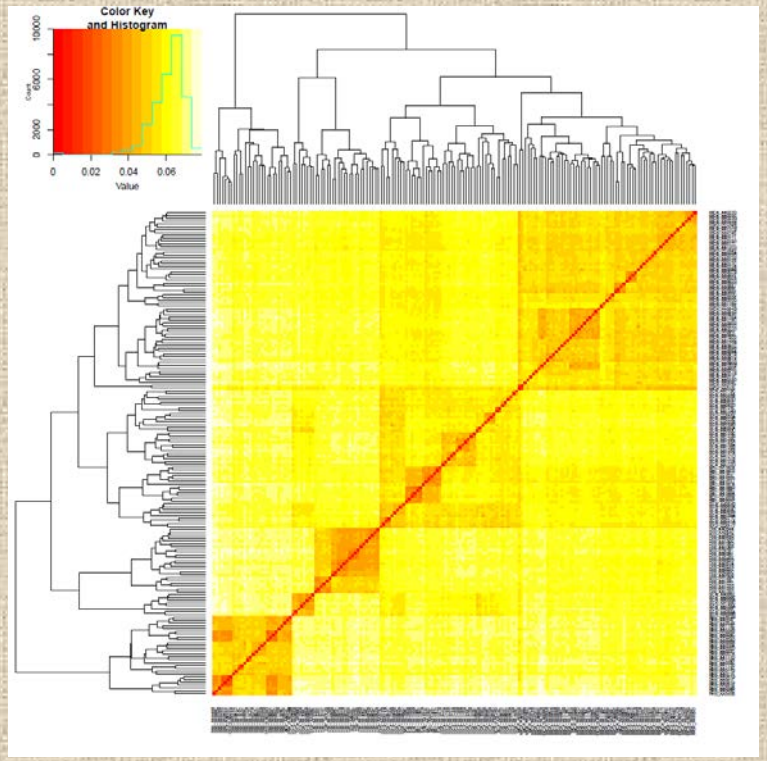
Analysis of single nucleotide polymorphism (SNP) data




```
2   25723   .   T   A   243640   PASS
AC=235;AF=0.996;AN=236;BaseQRankSum=0.736;ClippingRankSum=-0.736;DP=8928;FS=0;GQ_MEAN=142.44;GQ_STDDEV=
736;NCC=69;QD=32.98;ReadPosRankSum=-0.736;SOR=15.985 GT:AD:DP:GQ:PL  1/1:0,17:17:51:701,51,0 1/1:0,10:1
1/1:0,38:38:99:1587,114,0   ./.:115,0:115:.:.   1/1:0,9:9:27:373,27,0   ./.:21,0:21:.:. 1/1:0,4:4:12:16
1/1:0,7:7:21:282,21,0   ./.:107,0:107:.:.   1/1:0,13:13:39:534,39,0 ./.:6,0:6:.:.   1/1:0,28:28:84:1125
./.:11,0:11:.:. 1/1:0,38:38:99:1553,114,0   0/1:1,5:6:22:191,0,22   1/1:0,11:11:33:423,33,0 ./.:11,0:11
1/1:0,24:24:72:982,72,0 1/1:0,25:25:75:1036,75,0    ./.:4,0:4:.:.   ./.:39,0:39:.:. ./.:14,0:14:.:. 1/1
1/1:0,7:7:21:256,21,0   1/1:0,9:9:27:363,27,0   1/1:0,9:9:27:364,27,0   1/1:0,12:12:36:489,36,0 1/1:0,3
./.:82,0:82:.:. 1/1:0,68:68:99:2806,205,0   1/1:0,2:2:6:80,6,0  ./.:0,0:0:.:.   1/1:0,19:19:57:765,57,0
./.:0,0:0:.:.   1/1:0,28:28:84:1150,84,0    1/1:0,28:28:84:1161,84,0    ./.:0,0:0:.:.   ./.:0,0:0:.:.
1/1:0,40:40:99:1653,120,0   ./.:44,0:44:.:. ./.:99,0:99:.:. 1/1:0,15:15:45:621,45,0 ./.:27,0:27:.:. 1/1
1/1:0,27:27:81:1102,81,0    1/1:0,52:52:99:2115,157,0   1/1:0,81:81:99:3351,244,0   ./.:67,0:67:.:. 1/1
1/1:0,80:80:99:3294,241,0   1/1:0,10:10:30:410,30,0 ./.:52,0:52:.:. 1/1:0,36:36:99:1477,108,0   1/1:0,2
1/1:0,59:59:99:2431,178,0   1/1:0,30:30:90:1198,90,0    1/1:0,17:17:51:664,51,0 1/1:0,25:25:75:1028,75,
1/1:0,14:14:42:580,42,0 ./.:134,0:134:.:.   1/1:0,54:54:99:2215,163,0   ./.:39,0:39:.:. ./.:70,0:70:.:.
1/1:0,17:17:51:695,51,0 1/1:0,25:25:75:1030,75,0    ./.:14,0:14:.:. ./.:13,0:13:.:. ./.:15,0:15:.:. 1/1
1/1:0,21:21:63:859,63,0 ./.:32,0:32:.:. 1/1:0,26:26:78:1073,78,0    ./.:10,0:10:.:. 1/1:0,26:26:78:1076
./.:43,0:43:.:. 1/1:0,34:34:99:1399,102,0   ./.:0,0:0:.:.   1/1:0,58:58:99:2402,175,0   1/1:0,54:54:99:
1/1:0,34:34:99:1412,102,0   1/1:0,74:74:99:3077,223,0   1/1:0,116:116:99:4809,349,0 1/1:0,65:65:99:2679
1/1:0,93:93:99:3827,280,0   1/1:0,75:75:99:3110,226,0   1/1:0,38:38:99:1569,114,0   ./.:0,0:0:.:.   1/1
1/1:0,26:26:78:1080,78,0    1/1:0,41:41:99:1699,123,0   1/1:0,80:80:99:3323,241,0   1/1:0,85:85:99:3501
1/1:0,78:78:99:3223,235,0   ./.:109,0:109:.:.   1/1:0,138:138:99:5710,415,0 1/1:0,86:86:99:3540,259,0
1/1:0,49:49:99:2031,147,0   ./.:38,0:38:.:. 1/1:0,89:89:99:3684,268,0   1/1:0,142:142:99:5857,427,0 1/1
./.:57,0:57:.:. 1/1:0,12:12:36:492,36,0 1/1:0,83:83:99:3446,250,0   1/1:0,88:88:99:3633,265,0   1/1:0,6
1/1:0,32:32:96:1323,96,0    1/1:0,92:92:99:3813,277,0   1/1:0,44:44:99:1819,132,0   1/1:0,142:142:99:58
1/1:0,68:68:99:2801,205,0   1/1:0,113:113:99:4675,340,0 1/1:0,132:132:99:5484,397,0 ./.:108,0:108:.:.
2   25749   .   C   T   82850.5 PASS
AC=80;AF=0.253;AN=312;BaseQRankSum=-0.091;ClippingRankSum=-0.094;DP=9040;FS=0;GQ_MEAN=232;GQ_STDDEV=486
CC=22;QD=27.58;ReadPosRankSum=0.29;SOR=1.637 GT:AD:DP:GQ:PGT:PID:PL   0/1:3,14:17:84:.:.:546,0,84 1/1:0,
./.:3,0:3:.:.:.:.   0/0:39,0:39:0:.:.:0,0,480   0/0:115,0:115:0:.:.:0,0,401 0/1:5,4:9:99:.:.:146,0,194
1/1:0,5:5:15:1|1:25749_C_T:225,15,0 0/0:7,0:7:9:.:.:0,9,135 0/0:107,0:107:60:.:.:0,60,900   1/1:0,13:13
0/1:14,14:28:99:.:.:512,0,534   0/0:66,0:66:81:.:.:0,81,1215    1/1:0,13:13:39:1|1:25749_C_T:585,39,0
0/1:3,3:6:99:.:.:117,0,116  0/0:11,0:11:3:.:.:0,3,45    0/0:11,0:11:15:.:.:0,15,225 ./.:0,0:0:.:.:.:.
0/1:11,13:24:99:.:.:497,0,410   1/1:0,25:25:75:1|1:25749_C_T:1125,75,0  0/0:4,0:4:3:.:.:0,3,45  0/0:39,
0/0:93,0:93:93:.:.:0,93,1395    0/1:48,43:91:99:.:.:1618,0,1838 0/0:7,0:7:9:.:.:0,9,135 0/0:9,0:9:3:.:.
```

Filip Kolář

SNP

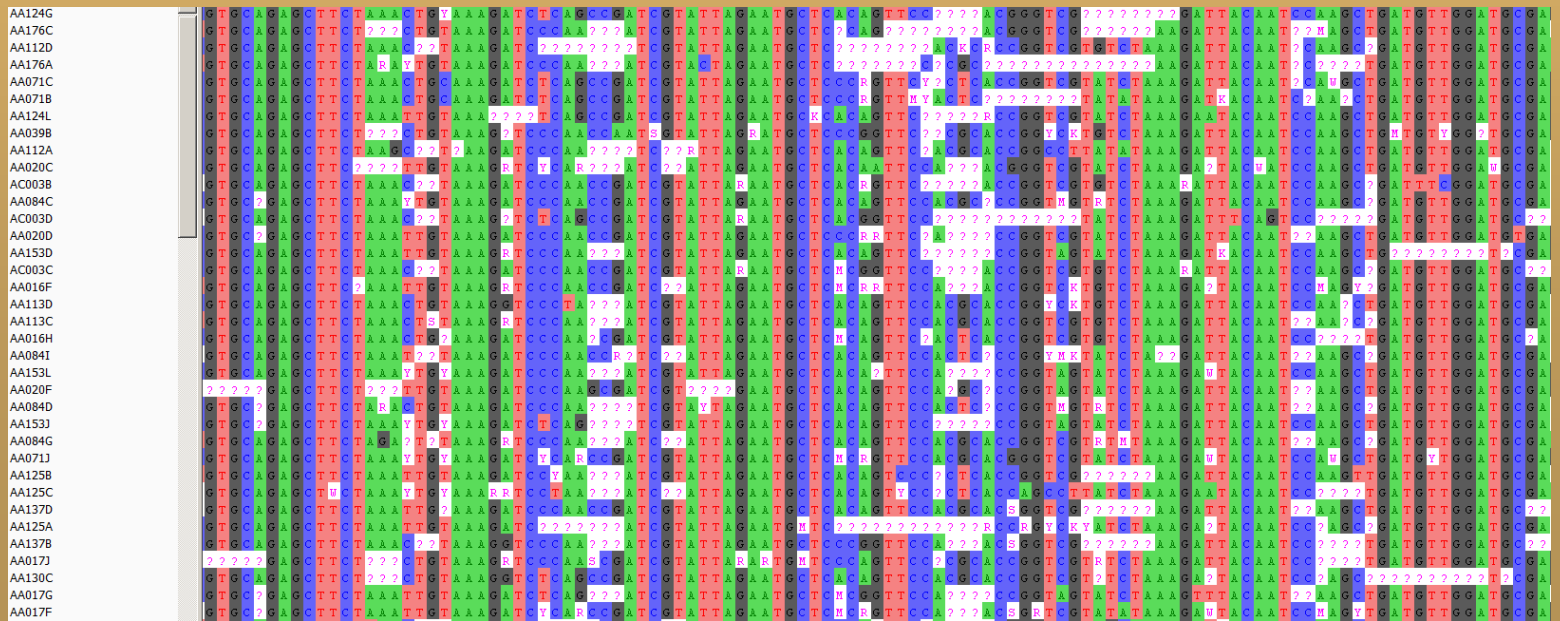
- single nucleotide polymorphisms
- max. four alleles (ATCG) - but usually biallelic

among 88 332 015 genetic variants identified in a sample of 2504 individuals, 95.53% were biallelic SNP, 4.07% insertions–deletions (indels), 0.33% multiallelic SNPs and 0.07% structural variants (The 1000 Genomes Project

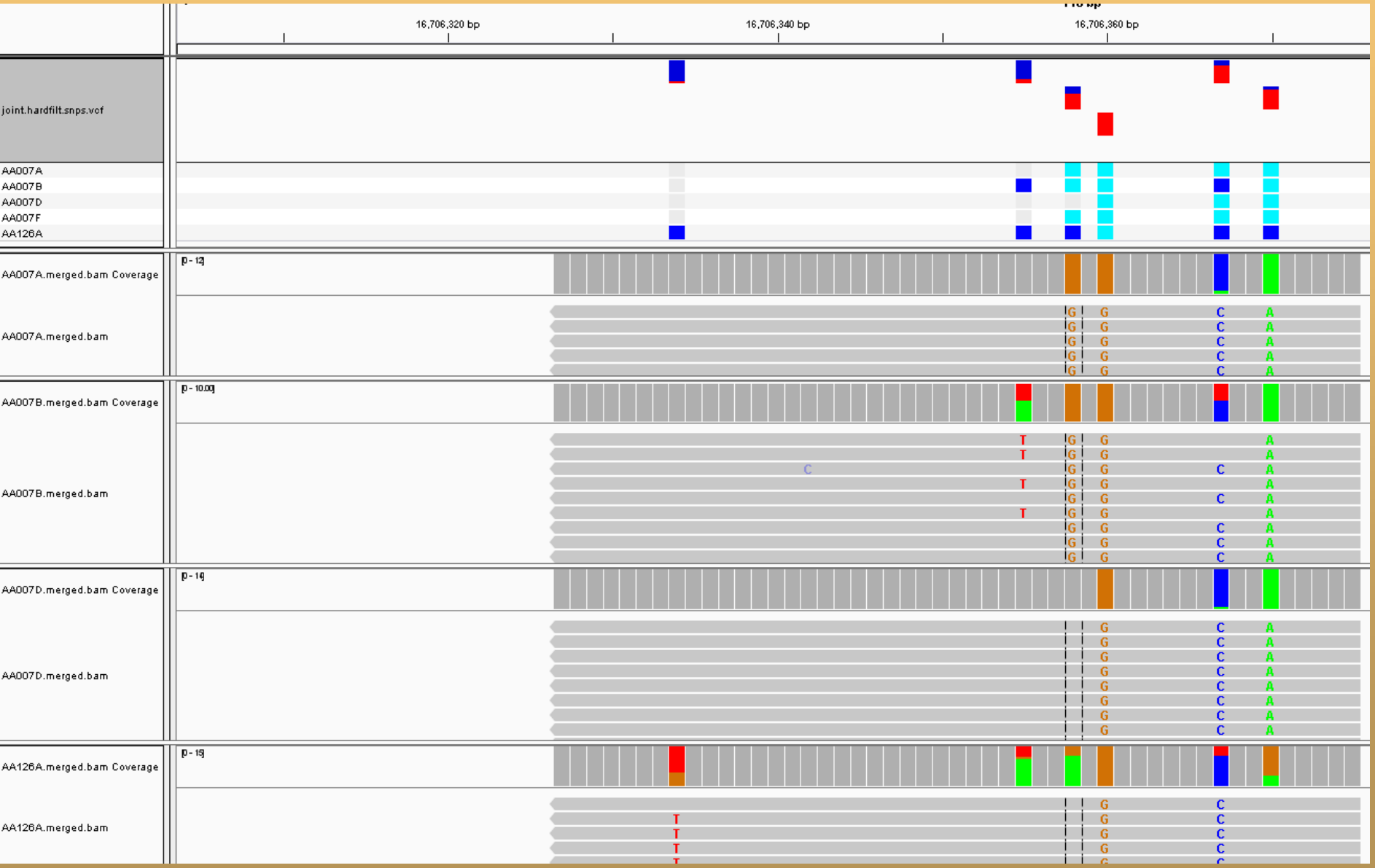
- codominant – homozygotes (e.g. AA, TT)
  x heterozygotes (e.g. AT)

- usually 1,000s – 10,000s (... up to tens of millions)

- substitution changes -> evolutionary models, coalescent simulations
- non-anonymous
- (un)linked ?!

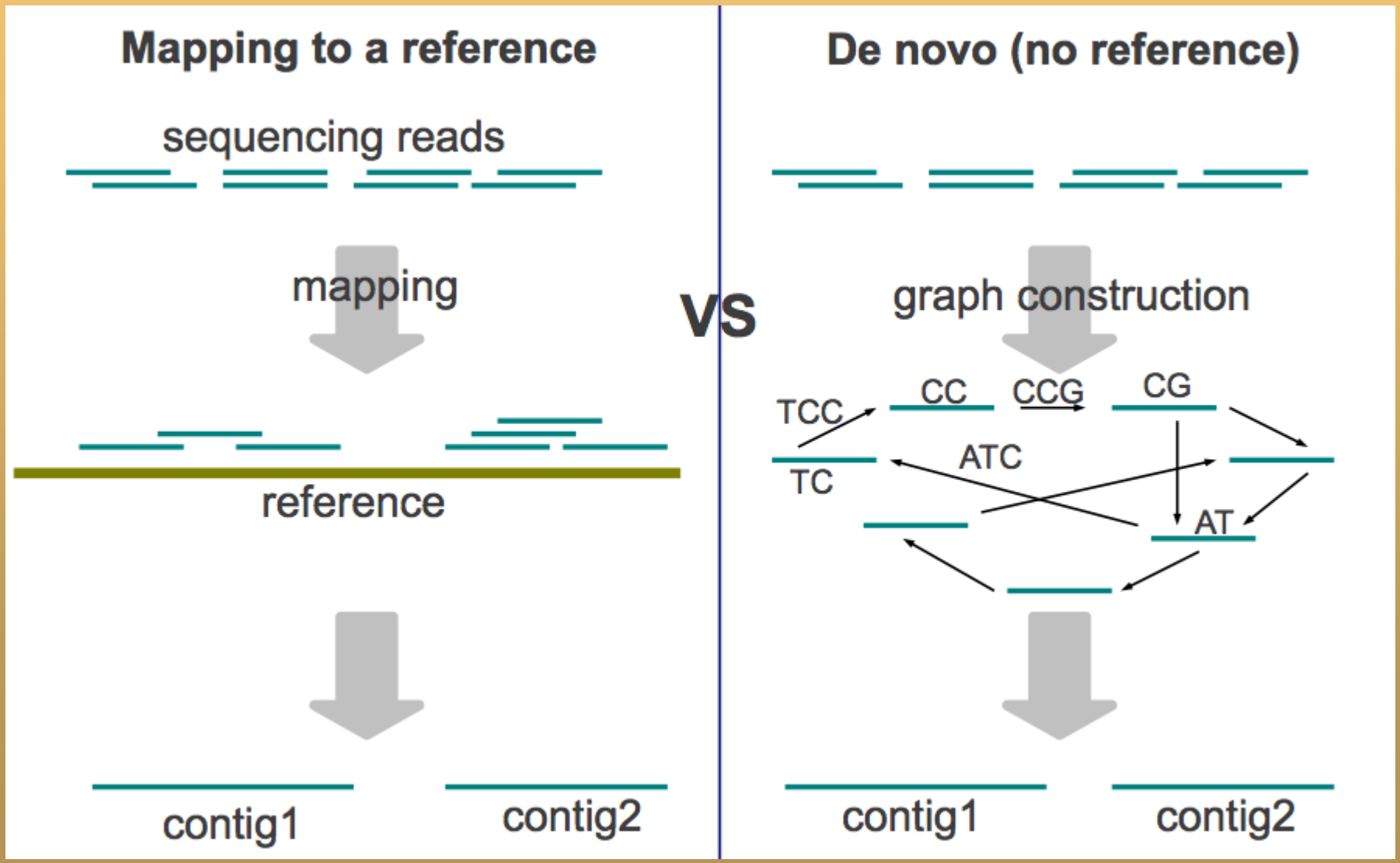


SNP


16,706,320 bp
16,706,340 bp
16,706,360 bp
joint.hardfilt.snps.vcf
AA007A
AA007B
AA007D
AA007F
AA126A
AA007A.merged.bam Coverage
AA007A.merged.bam
AA007B.merged.bam Coverage
AA007B.merged.bam
AA007D.merged.bam Coverage
AA007D.merged.bam
AA126A.merged.bam Coverage
AA126A.merged.bam

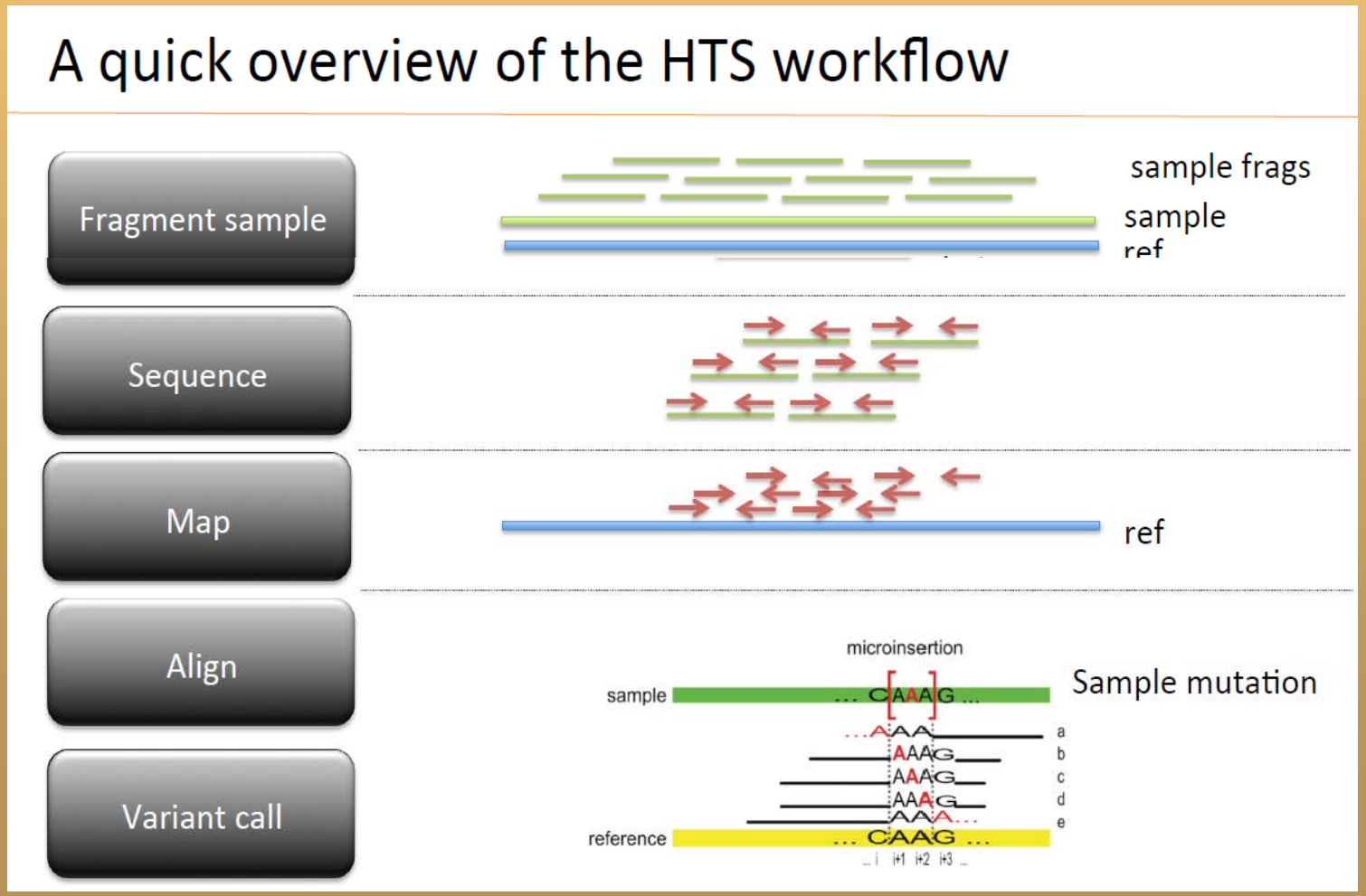
How to get to SNP data

- With or without reference



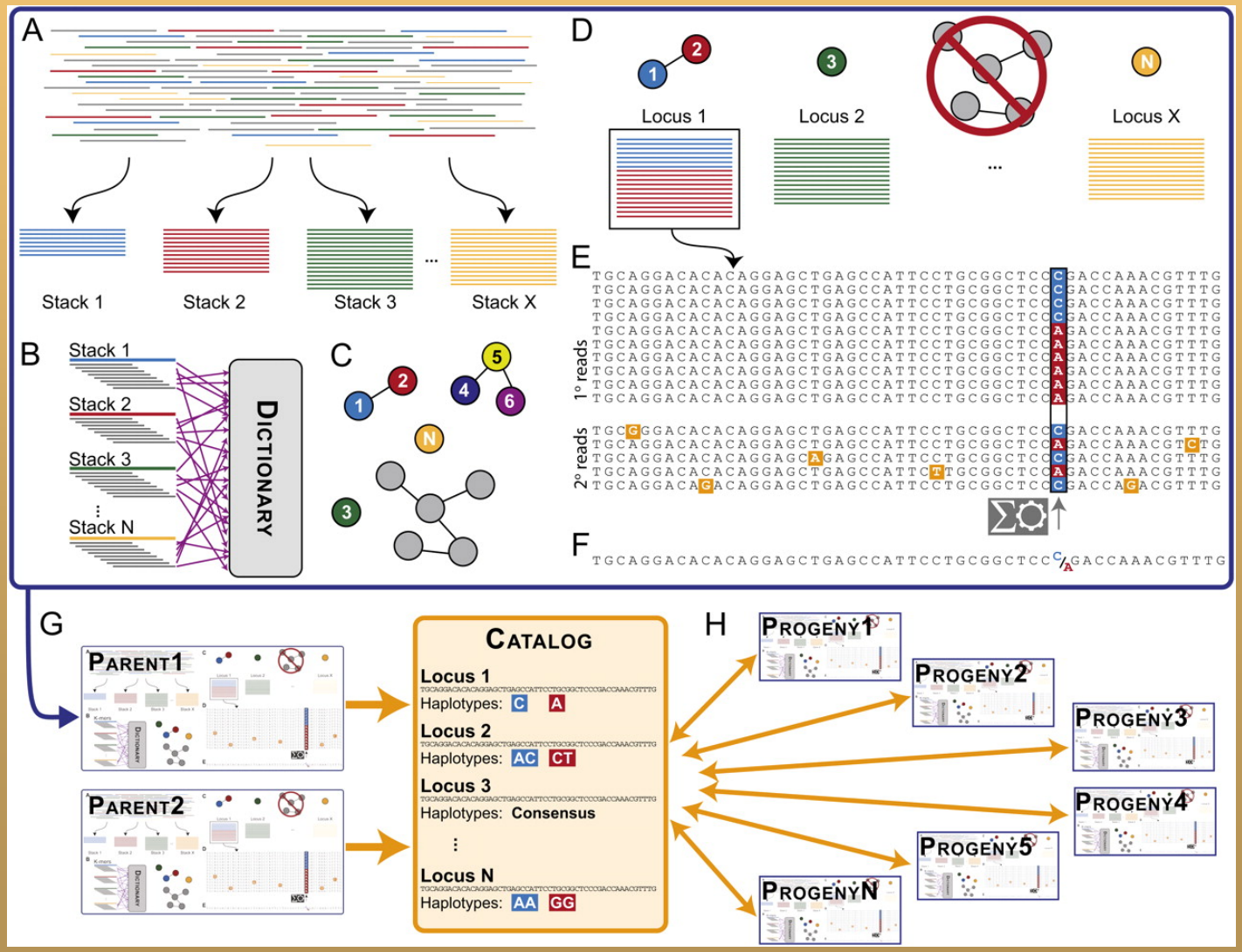
How to get to SNP data

- using high-throughput sequencing (HTS, NGS)
- A) With reference – whole genome resequencing / sub-sampling a genome (target enrichment / RADseq)



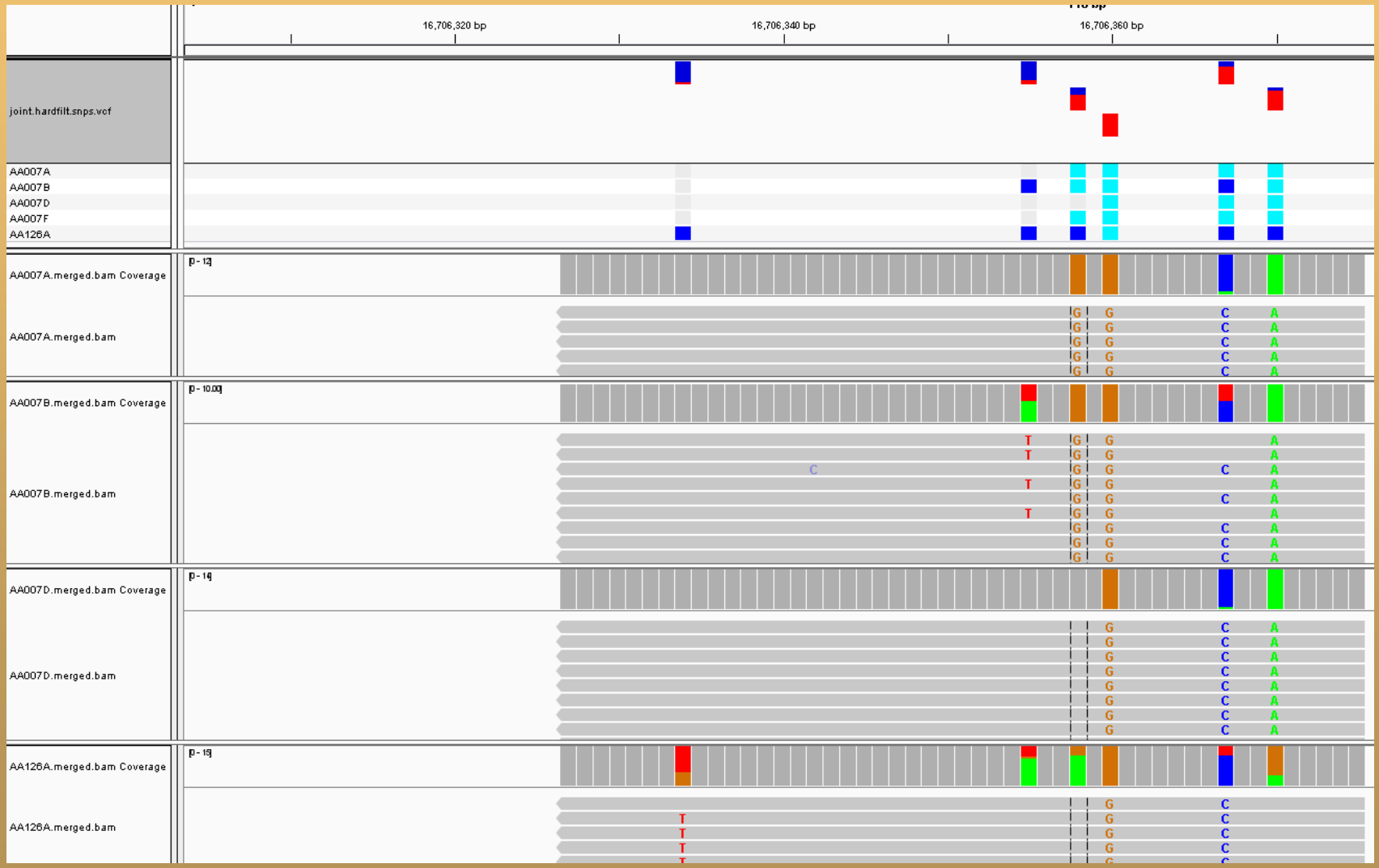
How to get to SNP data

- B) ... or without reference
- de-novo assembly of a “pseudo-reference” (e.g. RAD loci)



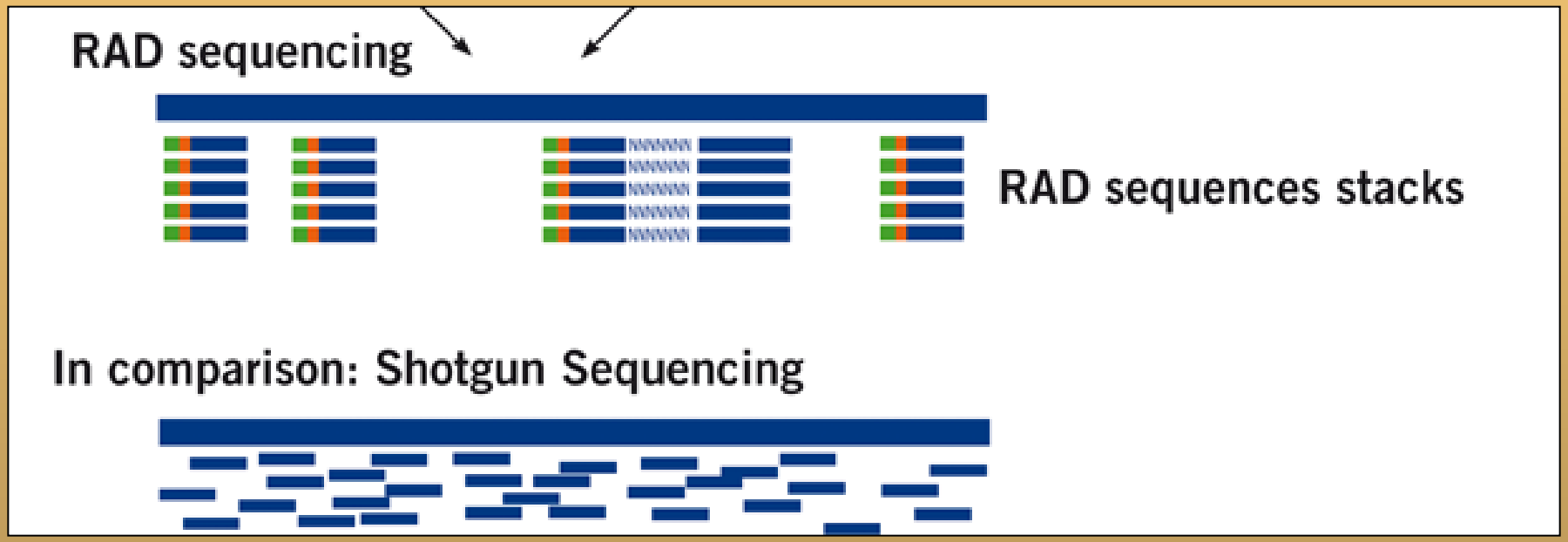
RAD sequencing

- restriction site associated DNA sequencing



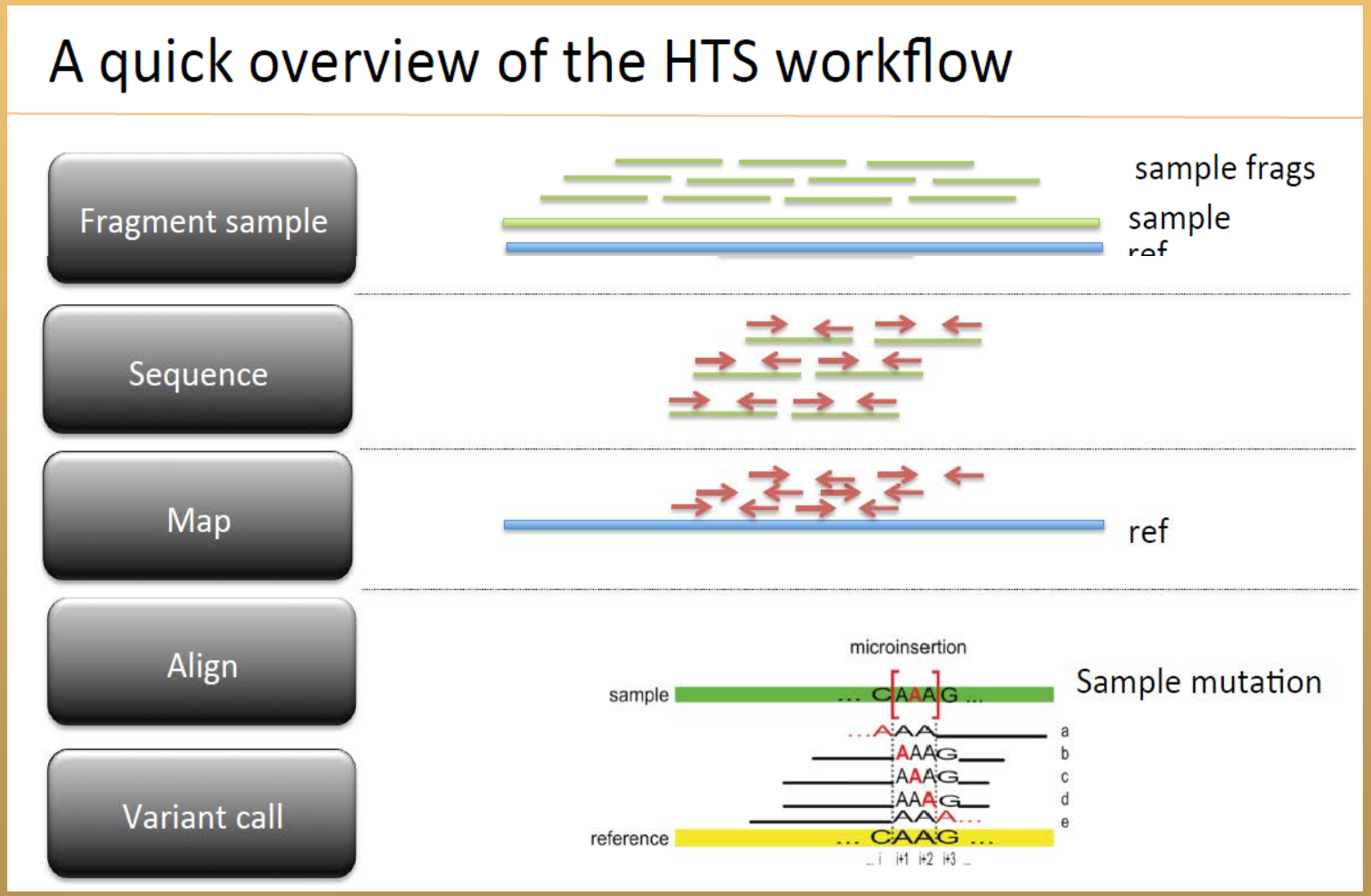
RAD sequencing

- restriction site associated DNA sequencing – pileup of reads
- double-digest RAD seq



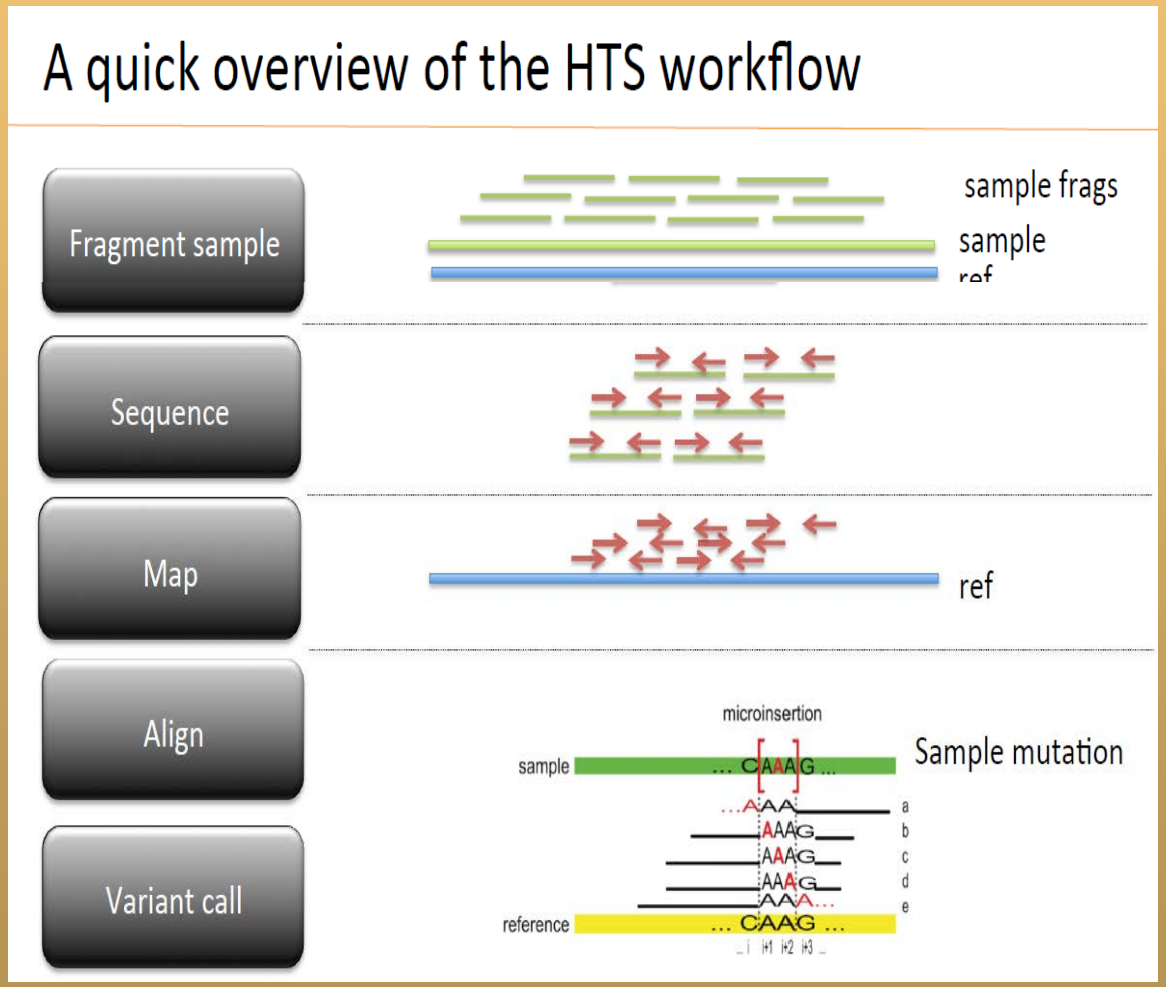
How to get to SNP data

- With reference



- ... and filter the variants / do variant recalibration

Files involved



- FASTQ
- SAM/BAM
- VCF

- … and filter the variants / do variant recalibration

Files involved

Genome (FASTA)

```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary
reference assembly
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
```

Reads (FASTQ)

```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

Mapped Reads (mpileup, BAM)

```
seq1 272 T 24  ,.$.....,.,.,.,.,.,.,..^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,.,.,.,.,.,.,..A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$.....,.,.,.,.,.,....    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$.....,.,.,.,.,.,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,.,.,.,.,.,...  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  .....,.,.,.C.,.,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  .....,.,.,.,.,.,.,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,.,.,.,.,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

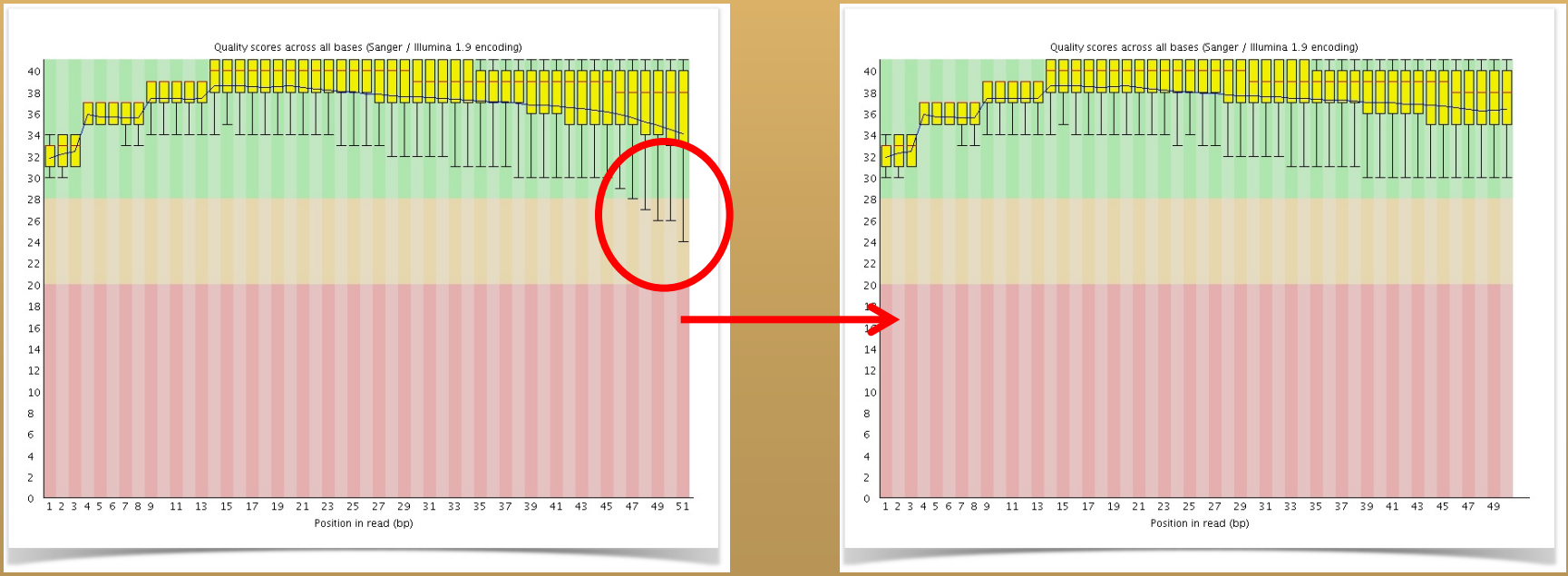
Variants (VCF)

```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
chr1    82154   rs4477212       a       .       .       .       GT      0
/0
chr1    752566  rs3094315       g       A       .       .       .       GT      1
/1
chr1    752721  rs3131972       A       G       .       .       .       GT      1
/1
chr1    798959  rs11240777      g       .       .       .       .       GT      0
/0
chr1    800007  rs6681049       T       C       .       .       .       GT      1
/1
```

Read trimming, Quality check

- *e.g. **Trimmomatic**, fastx toolkit*
- manipulating FASTQ files

- quality filtering
- quality trimming



FASTQ

- = fasta with qualities
- 1 read = 4 lines

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- p = the probability that the corresponding base call is wrong
- Qualities $Q_{\text{sanger}} = -10 \log_{10} p$
  - p = 0.1 ➔ Q = 10
  - p = 0.01 ➔ Q = 20
  - P = 0.001 ➔ Q = 30
- Encoding: Sanger/Phred format can encode a quality score from 0 to 93 using ASCII 33 to 126: Q + 33 ➔ ASCII code

```
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
 |                         |    |        |
33                        59   64       73
 0.........................26...31.......40
```

Read mapping

- *e.g. **BWA**, Bowtie, Stampy*
- unlike Sanger, we do not know the read's origin

(no primers, just "random" DNA fragments)

Millions of reads

Billions of positions in human genome

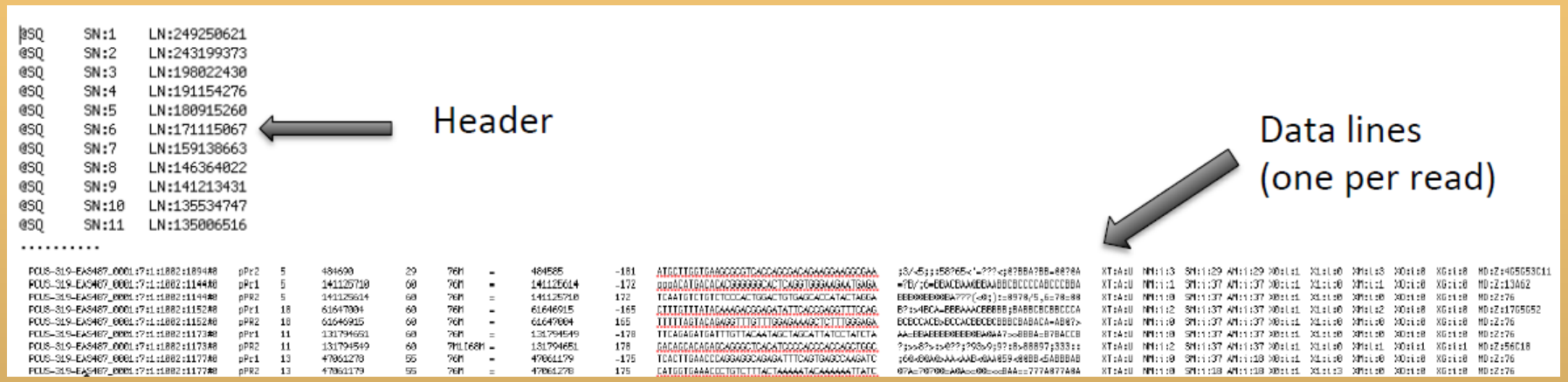
for **each read**:

- determine it's likely origin
- how likely it is we have correctly identified its origin
- not necessary to get exact alignmnet (later step – realignment around indels)

Reference	NNNNNCAAGNNNN	Reference	NNNNNCA_AGGNNN
Sample	NNNNNCAAAGNNN	Correct read align	NNNNNCAAAGNNNN
		Reference	NNNNNCAAGGNNN
		Alt. align	NNNNNCAAAGNNNN

SAM / BAM

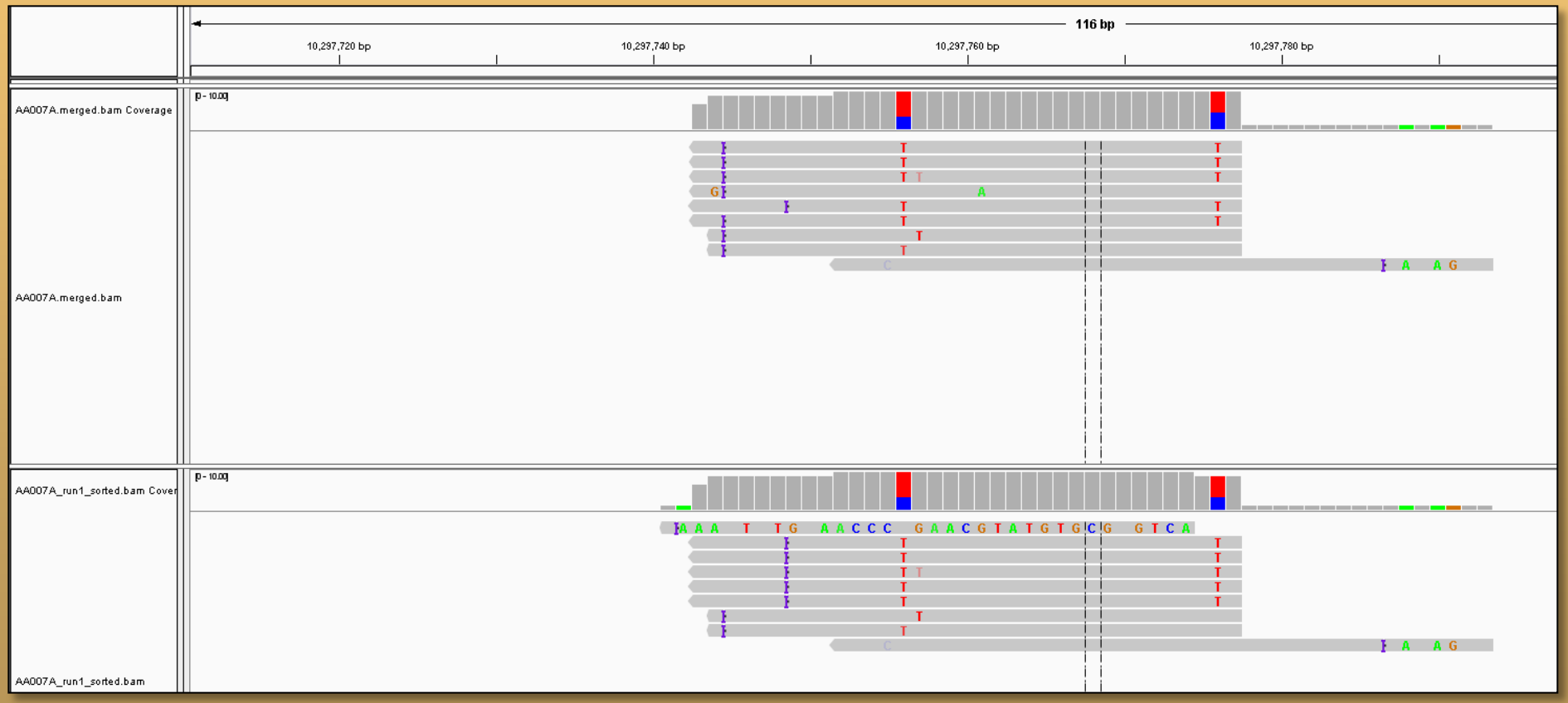
- reads mapped to a reference




```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

SAM / BAM

- reads mapped to a reference



- visualization in IGV – two individuals

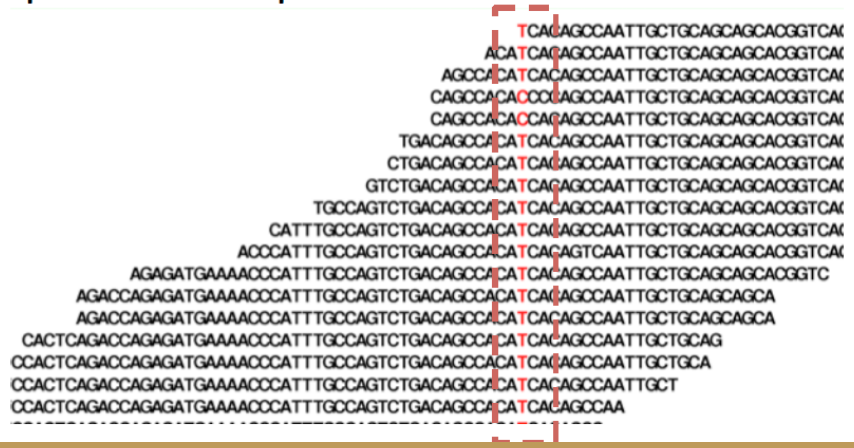
Variant calling

- *GATK, Samtools, FreeBayes*
- likelihood-based models - **Genotype likelihoods ->**

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

- **Sanger**: both alleles are amplified and sequenced at the same time
- **NGS**: each allele is sequenced separately and sampled with replacement



How many genotype likelihoods do we have for each individual at each site?

3 if both alleles are known
10 if not

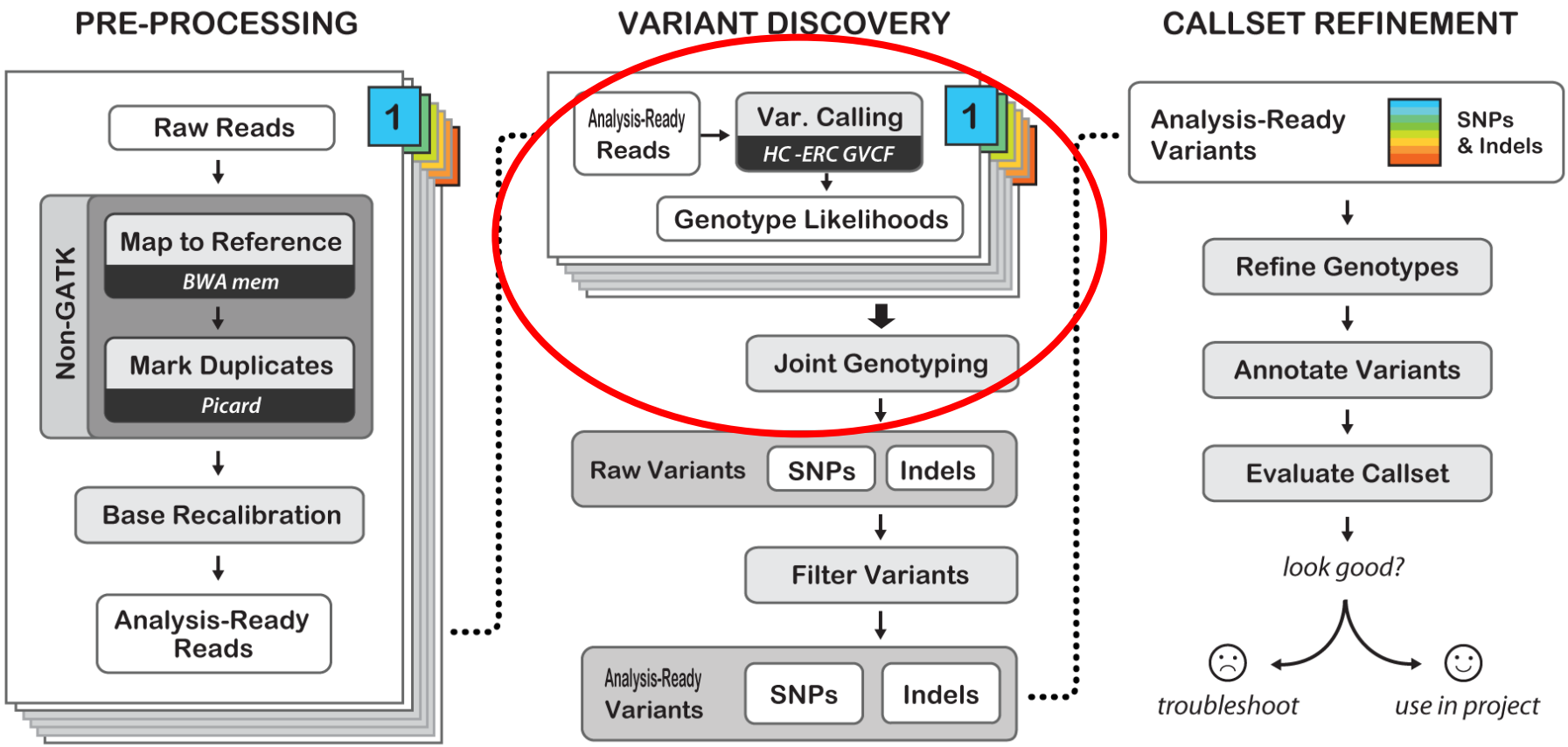
	A	C	G	T
A	1	2	3	4
C		5	6	7
G			8	9
T				10

- in any calling – large number of false positives and false negatives => filtering/variant recalibration

- paradigm: do not filter until VCF is produced, then apply **filtration**

Variant calling - GATK

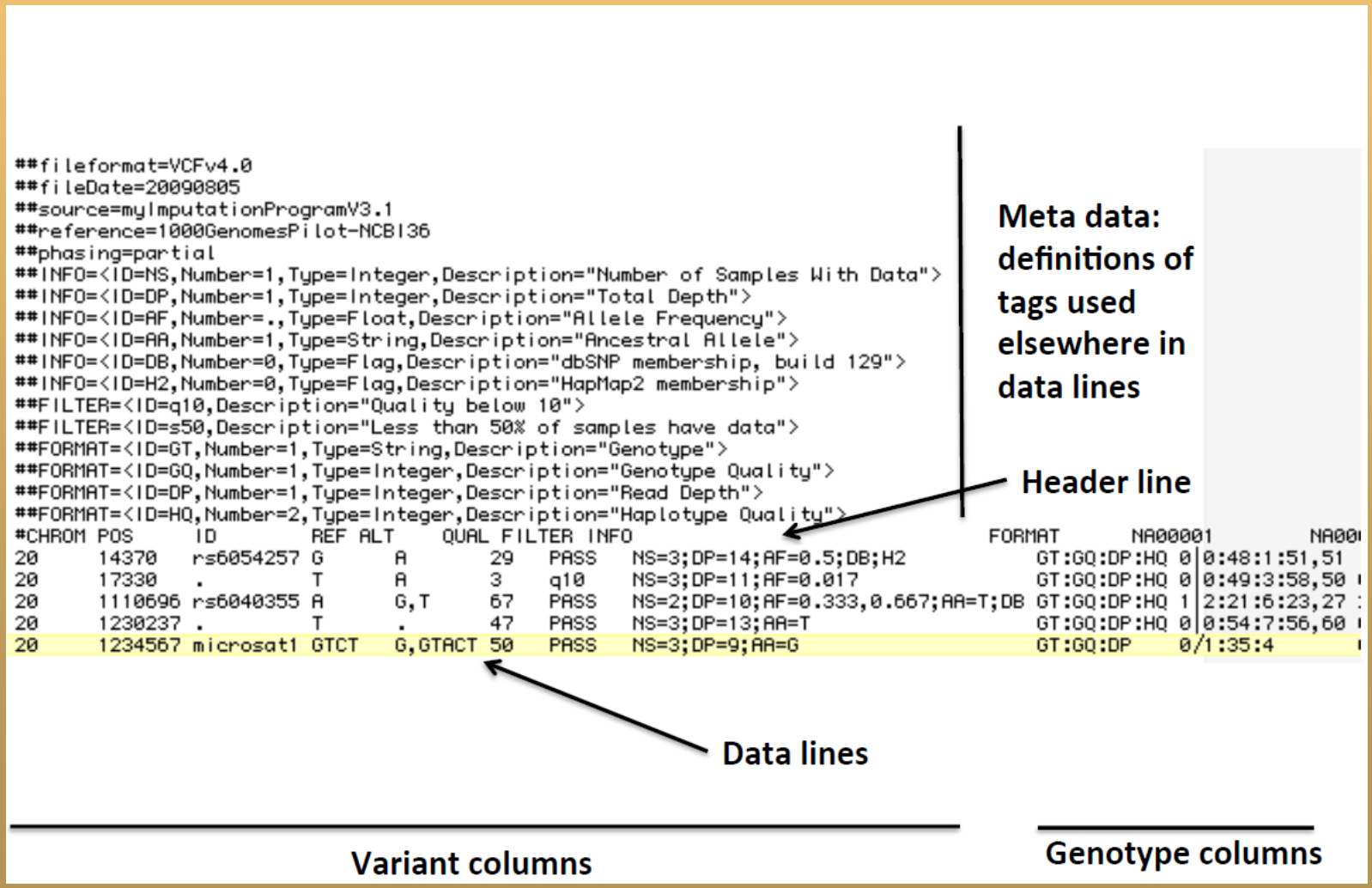
- *GATK best practice (human data)*
- https://software.broadinstitute.org/gatk/best-practices/bp_3step.php?case=GermShortWGS



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

VCF

- “golden standard” for SNP data
- SNPs = rows, columns = info on SNPs and individual genotypes



VCF

- “golden standard” for SNP data
- SNPs = rows, columns = info on SNPs and individual genotypes

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  WCA_AA007A  WCA_AA007B  WCA_AA007D  WCA_AA007F  WCA_AA009B  WCA_AA009I  PAN_AA011D  PAN_AA011E  WCA_AA012A  WCA_AA012C
WCA_AA012E  WCA_AA012H  WCA_AA013I  WCA_AA016F  WCA_AA016H  WCA_AA017F  WCA_AA017G  WCA_AA017I  WCA_AA017J  WCA_AA018D  WCA_AA018F
2   25723   .   T   A   243640  PASS
AC=235;AF=0.996;AN=236;BaseQRankSum=0.736;ClippingRankSum=-0.736;DP=8928;FS=0;GQ_MEAN=142.44;GQ_STDDEV=114.34;InbreedingCoeff=-0.0065;MLEAC=249;MLEAF=0.996;MQ=75.86;MQ0=0;MQRankSum=(
736;NCC=69;QD=32.98;ReadPosRankSum=-0.736;SOR=15.985 GT:AD:DP:GQ:PL  1/1:0,17:17:51:701,51,0 1/1:0,10:10:30:410,30,0 1/1:0,13:13:39:531,39,0 ./.:3,0:3:.:.
1/1:0,38:38:99:1587,114,0   ./.:115,0:115:.:.   1/1:0,9:9:27:373,27,0   ./.:21,0:21:.:. 1/1:0,4:4:12:163,12,0   1/1:0,8:8:24:330,24,0   1/1:0,5:5:15:205,15,0
1/1:0,7:7:21:282,21,0   ./.:107,0:107:.:.   1/1:0,13:13:39:534,39,0 ./.:6,0:6:.:.   1/1:0,28:28:84:1125,84,0   ./.:67,0:67:.:. 1/1:0,13:13:39:530,39,0 1/1:0,28:28:84:1150,84,0
./.:11,0:11:.:. 1/1:0,38:38:99:1553,114,0   0/1:1,5:6:22:191,0,22   1/1:0,11:11:33:423,33,0 ./.:11,0:11:.:. ./.:0,0:0:.:.   ./.:241,0:241:.:.
2   25749   .   C   T   82850.5 PASS
AC=80;AF=0.253;AN=312;BaseQRankSum=-0.091;ClippingRankSum=-0.094;DP=9040;FS=0;GQ_MEAN=232;GQ_STDDEV=486.67;InbreedingCoeff=0.3379;MLEAC=94;MLEAF=0.273;MQ=75.85;MQ0=0;MQRankSum=0.225
CC=22;QD=27.58;ReadPosRankSum=0.29;SOR=1.637 GT:AD:DP:GQ:PGT:PID:PL  0/1:3,14:17:84:.:.:546,0,84 1/1:0,10:10:30:1|1:25749_C_T:450,30,0   0/1:8,5:13:99:.:.:182,0,307
./.:3,0:3:.:.:.:.   0/0:39,0:39:0:.:.:0,0,480   0/0:115,0:115:0:.:.:0,0,401 0/1:5,4:9:99:.:.:146,0,194  0/0:21,0:21:24:.:.:0,24,360 0/0:5,0:5:3:.:.:0,3,45  0/0:8,0:8:6:.:.:0,6,90
1/1:0,5:5:15:1|1:25749_C_T:225,15,0 0/0:7,0:7:9:.:.:0,9,135 0/0:107,0:107:60:.:.:0,60,900   1/1:0,13:13:39:1|1:25749_C_T:585,39,0   0/0:6,0:6:6:.:.:0,6,90
0/1:14,14:28:99:.:.:512,0,534   0/0:66,0:66:81:.:.:0,81,1215    1/1:0,13:13:39:1|1:25749_C_T:585,39,0   0/1:13,15:28:99:.:.:571,0,497   ./.:11,0:11:.:.:.:. ./.:39,0:39:.:.:.:.
0/1:3,3:6:99:.:.:117,0,116  0/0:11,0:11:3:.:.:0,3,45    0/0:11,0:11:15:.:.:0,15,225 ./.:0,0:0:.:.:.:.
2   79982   .   G   T   9505.17 PASS
AC=2;AF=0.005464;AN=322;BaseQRankSum=1.63;ClippingRankSum=1.4;DP=22545;FS=0;GQ_MEAN=112.34;GQ_STDDEV=48.12;InbreedingCoeff=0.5032;MLEAC=2;MLEAF=0.005464;MQ=82.47;MQ0=0;MQRankSum=1.6
NCC=11;QD=32.38;ReadPosRankSum=-0.966;SOR=9.455  GT:AD:DP:GQ:PL  0/0:65,0:65:99:0,120,1800   0/0:36,0:36:99:0,108,1598   0/0:54,0:54:99:0,120,1800   0/0:24,0:24:72:0,72,1008
0/0:62,0:62:99:0,120,1800   0/0:250,0:250:99:0,120,1800 0/0:33,0:33:99:0,99,1448    ./.:0,0:0:.:.   0/0:15,0:15:0:0,0,587   0/0:61,0:61:99:0,120,1800   0/0:15,0:15:45:0,45,609
0/0:54,0:54:99:0,117,1800   0/0:180,0:180:99:0,120,1800 0/0:34,0:34:99:0,102,1404   0/0:42,0:42:99:0,112,1688   0/0:51,0:51:99:0,108,1800   0/0:60,0:60:99:0,120,1800
0/0:36,0:36:99:0,108,1490   ./.:0,0:0:.:.   0/0:18,0:18:54:0,54,731 0/0:47,0:47:99:0,120,1800   0/0:10,0:10:30:0,30,423 0/0:26,0:26:78:0,78,1068    0/0:1,0:1:3:0,3,39
./.:0,0:0:.:.   0/0:250,0:250:99:0,120,1800 0/0:252,0:252:99:0,120,1800 0/0:35,0:35:99:0,105,1552   0/0:70,0:70:99:0,120,1800   ./.:0,0:0:.:.   ./.:0,0:0:.:.
0/0:16,0:16:13:0,13,633 0/0:104,0:104:99:0,120,1800 0/0:218,0:218:99:0,120,1800 0/0:250,0:250:99:0,120,1800 ./.:58,0:58:.:.
2   79984   .   T   G   8026.39 PASS
AC=5;AF=0.016;AN=328;BaseQRankSum=1.34;ClippingRankSum=-0.033;DP=22884;FS=0;GQ_MEAN=141.66;GQ_STDDEV=317.84;InbreedingCoeff=-0.019;MLEAC=6;MLEAF=0.016;MQ=81.04;MQ0=0;MQRankSum=0.214
CC=7;QD=15.35;ReadPosRankSum=0.917;SOR=0.435 GT:AD:DP:GQ:PL  0/0:65,0:65:99:0,120,1800   0/0:36,0:36:99:0,108,1598   0/0:54,0:54:99:0,120,1800   0/0:24,0:24:72:0,72,1008
0/1:50,30:80:99:1067,0,1908 0/1:119,160:282:99:6169,0,4328  0/0:33,0:33:99:0,99,1448    ./.:0,0:0:.:.   0/0:16,0:16:13:0,13,645 0/0:61,0:61:99:0,120,1800   0/0:15,0:15:45:0,45,609
0/0:54,0:54:99:0,117,1800   0/0:180,0:180:99:0,120,1800 0/0:34,0:34:99:0,102,1404   0/0:42,0:42:99:0,112,1688   0/0:51,0:51:99:0,108,1800   0/0:60,0:60:99:0,120,1800
0/0:36,0:36:99:0,108,1490   ./.:0,0:0:.:.   0/0:18,0:18:54:0,54,731 0/0:47,0:47:99:0,120,1800   0/0:10,0:10:30:0,30,423 0/0:26,0:26:78:0,78,1068    0/0:1,0:1:3:0,3,39
./.:0,0:0:.:.   0/0:250,0:250:99:0,120,1800 0/0:252,0:252:99:0,120,1800 0/0:35,0:35:99:0,105,1552   0/0:70,0:70:99:0,120,1800   ./.:0,0:0:.:.   ./.:0,0:0:.:.
0/0:16,0:16:2:0,2,638   0/0:104,0:104:99:0,120,1800 0/0:218,0:218:99:0,120,1800 0/0:250,0:250:99:0,120,1800
2   79986   .   G   A   10995.8 PASS
AC=2;AF=0.011;AN=326;BaseQRankSum=0.319;ClippingRankSum=1.45;DP=22821;FS=0;GQ_MEAN=139.02;GQ_STDDEV=263.84;InbreedingCoeff=-0.0179;MLEAC=4;MLEAF=0.011;MQ=82.19;MQ0=0;MQRankSum=4.74;
C=8;QD=18.51;ReadPosRankSum=0.811;SOR=0.48   GT:AD:DP:GQ:PGT:PID:PL  0/0:65,0:65:99:.:.:0,120,1800   0/0:36,0:36:63:.:.:0,63,1507    0/0:54,0:54:99:.:.:0,120,1800
0/0:24,0:24:72:.:.:0,72,1008    0/0:119,0:119:99:.:.:0,120,1800 0/0:323,0:323:99:.:.:0,120,1800 0/0:33,0:33:99:.:.:0,99,1448    ./.:0,0:0:.:.:.:.   0/0:16,0:16:0:.:.:0,0,555
0/0:61,0:61:99:.:.:0,120,1800   0/0:15,0:15:45:.:.:0,45,609 0/0:54,0:54:99:.:.:0,117,1800   0/0:180,0:180:99:.:.:0,120,1800 0/0:34,0:34:99:.:.:0,102,1404
0/0:42,0:42:99:.:.:0,112,1688   0/0:51,0:51:99:.:.:0,108,1800   0/0:60,0:60:99:.:.:0,120,1800   0/0:36,0:36:99:.:.:0,108,1490   ./.:0,0:0:.:.:.:.   0/0:18,0:18:54:.:.:0,54,731
0/0:47,0:47:99:.:.:0,120,1800   0/0:10,0:10:30:.:.:0,30,423 0/0:26,0:26:78:.:.:0,78,1068    0/0:1,0:1:3:.:.:0,3,39
```

VCF

- “golden standard” for SNP data
- SNPs = rows, columns = info on SNPs and individual genotypes

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	WCA_AA007A	WCA_AA007B	WCA_AA007D	WCA_AA007F	WCA_AA009B	WCA_AA00
2	25723	.	T	A	243640	PASS	.	GT:AD:DP:GQ:PL	1/1:0,17:17:51:701,51,0	1/1:0,10:10:30:410,30,0	1/1:0,13:13:39:5			
2	25749	.	C	T	82850.5	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	0/1:3,14:17:84:.:.:546,0,84	1/1:0,10:10:30:1|1:25749_C_T				
2	79982	.	G	T	9505.17	PASS	.	GT:AD:DP:GQ:PL	0/0:65,0:65:99:0,120,1800	0/0:36,0:36:99:0,108,1598	0/0:54,0			
2	79984	.	T	G	8026.39	PASS	.	GT:AD:DP:GQ:PL	0/0:65,0:65:99:0,120,1800	0/0:36,0:36:99:0,108,1598	0/0:54,0			
2	79986	.	G	A	10995.8	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	0/0:65,0:65:99:.:.:0,120,1800	0/0:36,0:36:63:.:.:0,63,				
2	79992	.	C	A	6577.46	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	0/0:65,0:65:99:.:.:0,120,1800	0/0:36,0:36:99:.:.:0,108				
2	80020	.	C	A	870517	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	1/1:0,63:63:99:.:.:2728,190,0	1/1:0,33:33:99:.:.:1441,				
2	216783	.	G	C	1639.73	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	0/0:21,0:21:63:.:.:0,63,845	0/0:16,0:16:48:.:.:0,48,643				
2	216790	.	A	G	11217	PASS	.	GT:AD:DP:GQ:PGT:PID:PL	./.:21,0:21:.:.:.:.	1/1:0,1:1:3:1|1:216786_G_A:45,3,0				

VCF

- here: first two SNPs for two individuals

```
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  WCA_AA007A  WCA_AA007B
2   25723   .   T   A   243640  PASS    AC=235;AF=0.996;AN=236;DP=8928  GT:AD:DP:GQ:PL  1/1:0,17:17:51:701,51,0 1/1:0,10:10:30:410,30,0
2   25749   .   C   T   82850.5 PASS    AC=80;AF=0.253;AN=312;DP=9040   GT:AD:DP:GQ:PGT:PID:PL  0/1:3,14:17:84:.:.:546,0,84 1/1:0,10:10:30:1|1:25749_C_T:450,30,0
```

1 2 3 4 5 6 7 8 9 10 11

- 1 – chromosome/scaffold
- 2 – position at chr./scaff.
- 3 ... often not used
- 4 – reference base
- 5 – alternative (non-ref) base
- 6 – SNP quality score (is the site variant?)
- 7 – filter field (PASS or filter name)
- 8 – “INFO field” – info on the SNP over all samples
- 9 – “FORMAT field” – definition of fields in next colum
- 10, ... - info on genotypes of each individual

- **CHROMO:** chromosome / contig
- **POS:** the reference position with the 1st base having position 1
- **ID:** an id; rs number if dbSNP variant
- **REF:** reference base.
  - The value in POS refers to the position of the first base in the string
  - for indels, the reference string must include the base before the event (and this must be reflected in POS)
- **ALT:** comma sepearated list of alternate non-ref alleles called on at least one of the samples
  - if no alternate alleles then the missing value should be used “.”
- **QUAL:** phred-scaled quality score of the assertion made in ALT (whether variant or non-variant)
- **FILTER:** PASS if the position has passed all filters (defined in meta-data).
- **INFO:** additional information

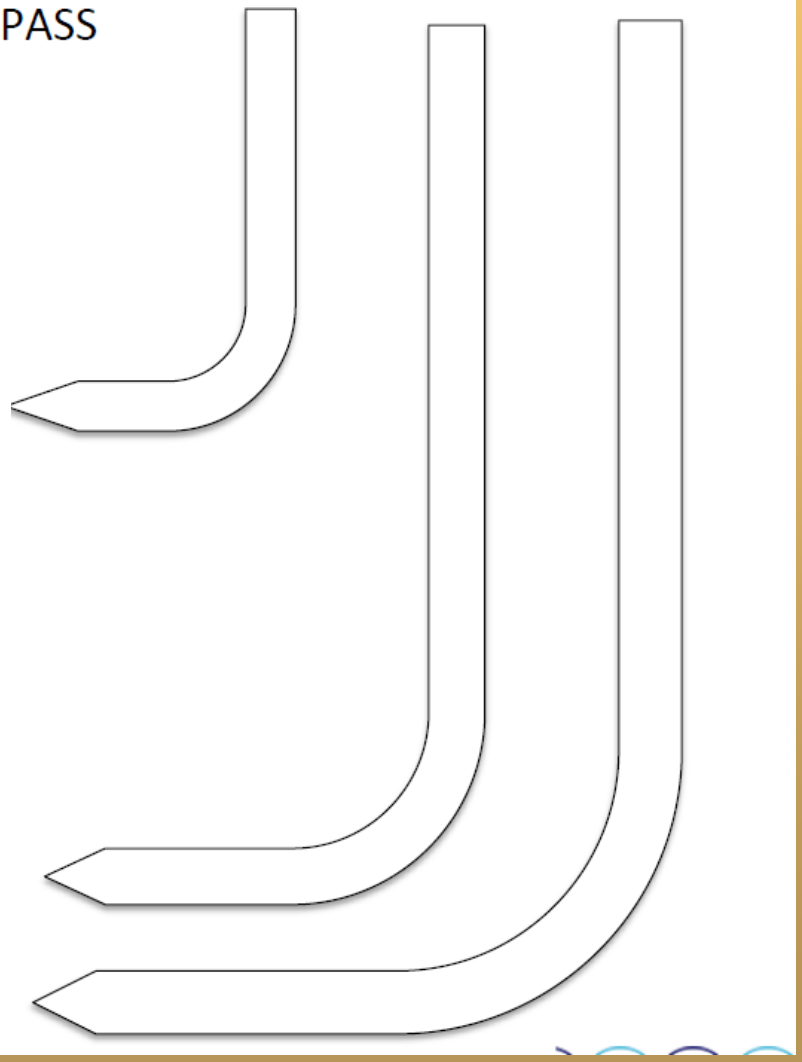
VCF

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample1
1 801943 rs7516866 C T 9787.34 PASS

AC=2;
AF=1.00;
AN=2;
BaseQRankSum=1.009;
DB;
DP=556;
FS=18.302;
MQ=44.04;
MQ0=38;
MQRankSum=5.122;
QD=17.60;
ReadPosRankSum=3.375

GT:AD:DP:GQ:PL

1/1:37,518:556:99:9787,685,0



VCF

- genotypes

C	T	GT:AD:DP:GQ:PL	1/1:0,17:17:51:701,51,0	1/1:0,10:10:30:410,30,0

- 1/1 = homozygote ALT
- 1/0 = heterozygote
- 0/0 = homozygote REF
- ./. = missing data

- Format field specifies type of data present for each genotype
  - GT:AD:DP:GQ:PL
  - fields defined in metadata header
- GT: genotype, encoded as alleles separated by either | or /
  - 0 for the ref, 1 for the 1st allele listed in ALT, 2 for the second, etc
  - REF=A and ALT=T
    - genotype 0/1 means hetero A/T
    - genotype 1/1 means homo T/T
  - /: genotype unphased and | genotype phased
- DP: read depth at position for sample
- GQ: genotype quality encoded as a phred quality
- etc.....

VCF – not only SNPs !!

- N. B. in an unfiltered VCF you may also see (not topic for today)

multiallelic SNPs

```
scaffold_1   128556   .   C   G,T     3139.27       PASS
```

insertions

```
scaffold_1   128556   .   G   GGGACCCT    3139.27       PASS
```

deletions

```
scaffold_1   128556   .   CTG   C     3139.27       PASS
```

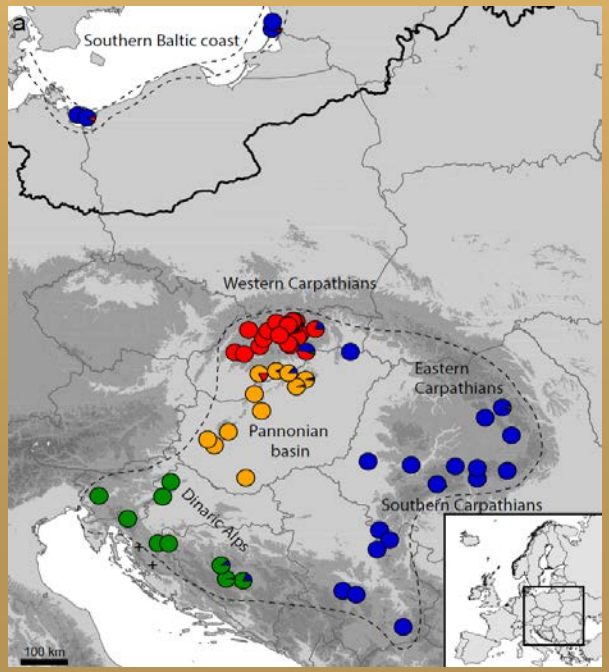
phased genotypes

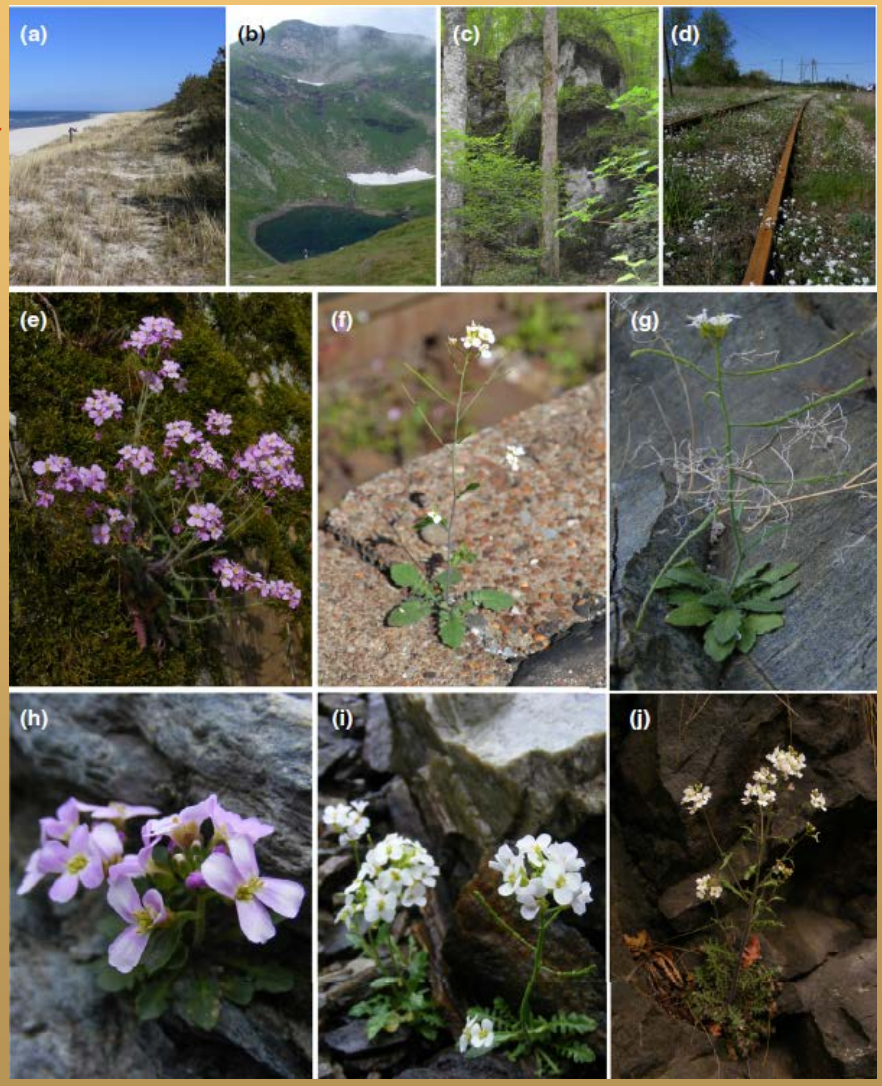
```
scaffold_1   128556   .   C   G     3139.27       PASS .
GT:AD:DP:GQ:PL  1|0:0,34:34:99:1412,102,0  1|1:0,74:74:99:3077,223,0
```

Today's tasks

- understand structure of the VCF
- explore the SNPs in entire VCF
- visualize genetic structure of a real dataset of a diploid plant in a multi-sample VCF (171 indivs from 64 pops, ~ 10,000 SNPs)
  - PCA, K-means clustering, distances, AMOVA, isolation by distance

- ... and answer the **questions** on the way

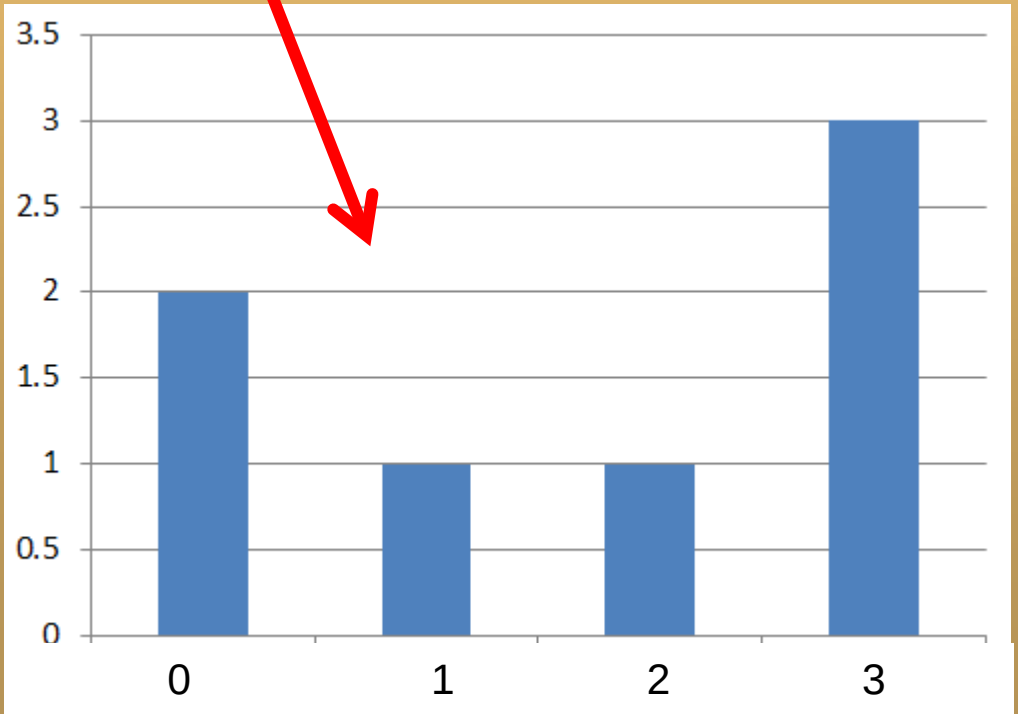




Arabidopsis arenosa

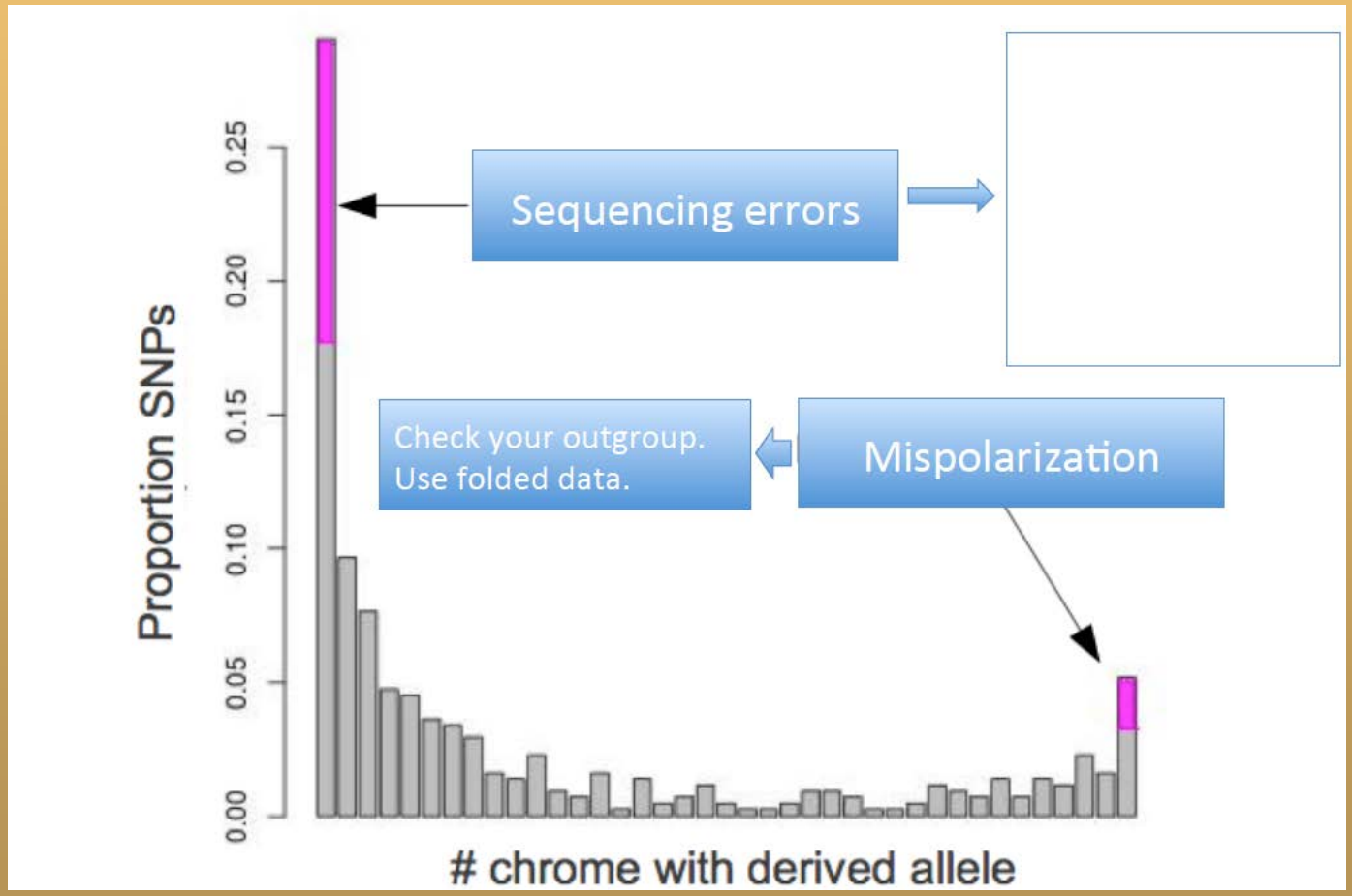
Allele frequency spectrum

```
                    123456
HumanSequence1 AATAGC
HumanSequence2 ..AC..
HumanSequence3 .TACT.
HumanSequence4 ..ACT.
```



Allele frequency spectrum

- effect of errors



Allele frequency spectrum

- biological interpretations