BIOINFORMATICS IN POPULATION GENETICS
BIN784

Gulsah Merve Kilinc, PhD
Lecturer@Department of Bioinformatics
Graduate School of Health Sciences
Hacettepe University

https://www.gulsahmervekilinc.com
email: gulsahkilinc@hacettepe.edu.tr & gulsahhdal@gmail.com

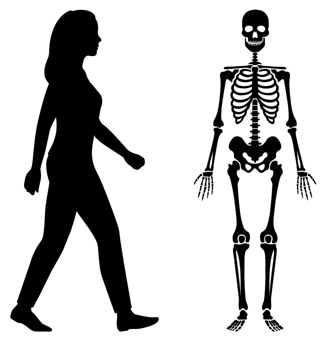
Course Materials

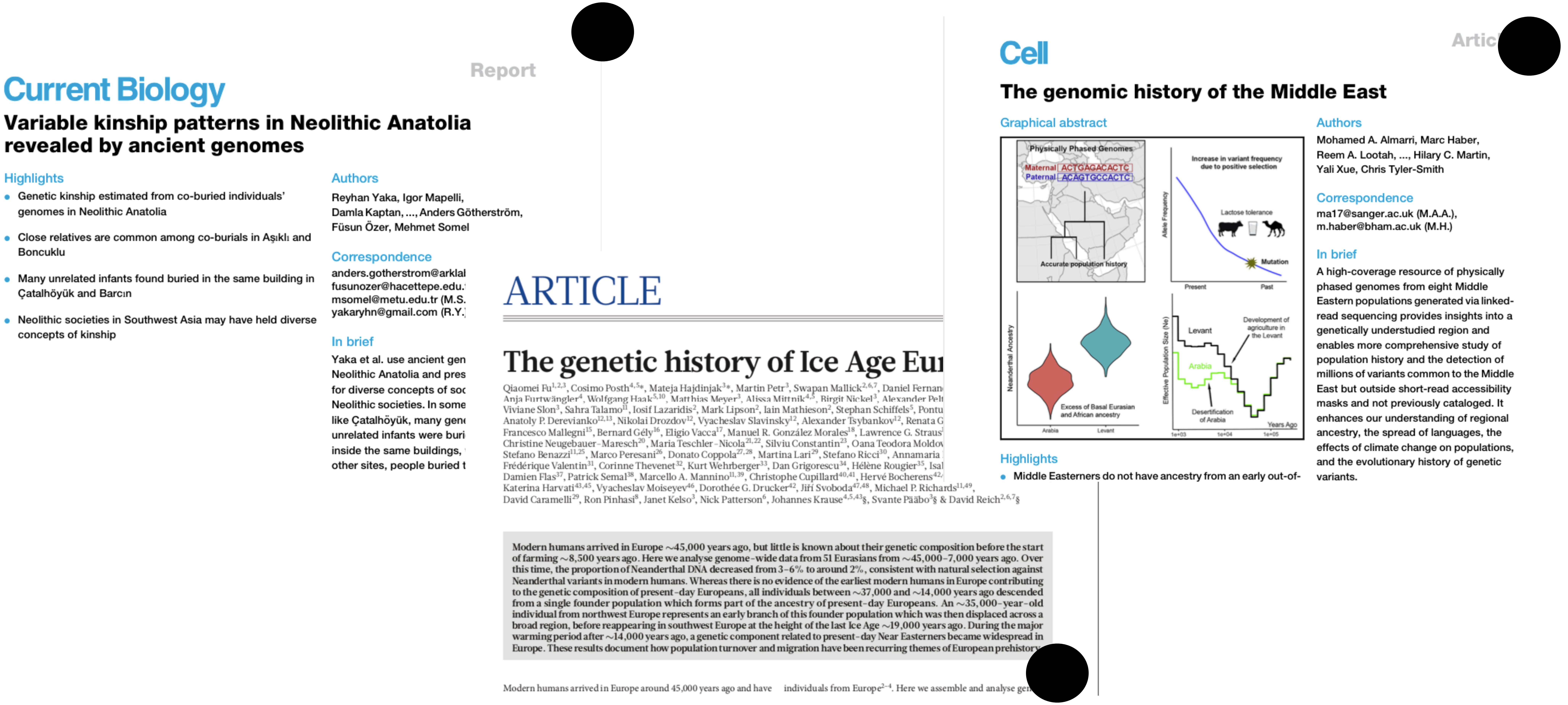
https://github.com/gulki/BIN784

- Nielsen, Rasmus; Slatkin, Montgomery. An introduction to population genetics : theory and applications, Sunderland, Mass.: Sinauer Associates, c2013
- Hamilton M, Population Genetics, Wiley-Blackwell
- Recently published papers

- Background in Linux, R and genome analysis is required

We will use published genome sequences during the course for learning- mostly human and/or ancient





Report

Current Biology

Variable kinship patterns in Neolithic Anatolia revealed by ancient genomes

Highlights

- Genetic kinship estimated from co-buried individuals' genomes in Neolithic Anatolia
- Close relatives are common among co-burials in Aşıklı and Boncuklu
- Many unrelated infants found buried in the same building in Çatalhöyük and Barcın
- Neolithic societies in Southwest Asia may have held diverse concepts of kinship

Authors

Reyhan Yaka, Igor Mapelli, Damla Kaptan, ..., Anders Götherström, Füsun Özer, Mehmet Somel

Correspondence

anders.gotherstrom@arklal
fusunozer@hacettepe.edu.
msomel@metu.edu.tr (M.S.
yakaryhn@gmail.com (R.Y.

In brief

Yaka et al. use ancient gen
Neolithic Anatolia and pres
for diverse concepts of soc
Neolithic societies. In some
like Çatalhöyük, many gene
unrelated infants were buri
inside the same buildings,
other sites, people buried t

ARTICLE

The genetic history of Ice Age Eur

Qiaomei Fu[1,2,3], Cosimo Posth[4,5]*, Mateja Hajdinjak[3]*, Martin Petr[3], Swapan Mallick[2,6,7], Daniel Fernan
Anja Furtwängler[4], Wolfgang Haak[5,10], Matthias Meyer[3], Alissa Mittnik[4,5], Birgit Nickel[3], Alexander Pelt
Viviane Slon[3], Sahra Talamo[11], Iosif Lazaridis[2], Mark Lipson[2], Iain Mathieson[2], Stephan Schiffels[5], Pontu
Anatoly P. Derevianko[12,13], Nikolai Drozdov[12], Vyacheslav Slavinsky[12], Alexander Tsybankov[12], Renata G
Francesco Mallegni[15], Bernard Gély[16], Eligio Vacca[17], Manuel R. González Morales[18], Lawrence G. Straus
Christine Neugebauer-Maresch[20], Maria Teschler-Nicola[21,22], Silviu Constantin[23], Oana Teodora Moldov
Stefano Benazzi[11,25], Marco Peresani[26], Donato Coppola[27,28], Martina Lari[29], Stefano Ricci[30], Annamaria
Frédérique Valentin[31], Corinne Thevenet[32], Kurt Wehrberger[33], Dan Grigorescu[34], Hélène Rougier[35], Isal
Damien Flas[37], Patrick Semal[38], Marcello A. Mannino[11,39], Christophe Cupillard[40,41], Hervé Bocherens[42,
Katerina Harvati[43,45], Vyacheslav Moiseyev[46], Dorothée G. Drucker[42], Jiří Svoboda[47,48], Michael P. Richards[11,49],
David Caramelli[29], Ron Pinhasi[8], Janet Kelso[3], Nick Patterson[6], Johannes Krause[4,5,43]§, Svante Pääbo[3]§ & David Reich[2,6,7]§

Modern humans arrived in Europe ~45,000 years ago, but little is known about their genetic composition before the start of farming ~8,500 years ago. Here we analyse genome-wide data from 51 Eurasians from ~45,000–7,000 years ago. Over this time, the proportion of Neanderthal DNA decreased from 3–6% to around 2%, consistent with natural selection against Neanderthal variants in modern humans. Whereas there is no evidence of the earliest modern humans in Europe contributing to the genetic composition of present-day Europeans, all individuals between ~37,000 and ~14,000 years ago descended from a single founder population which forms part of the ancestry of present-day Europeans. An ~35,000-year-old individual from northwest Europe represents an early branch of this founder population which was then displaced across a broad region, before reappearing in southwest Europe at the height of the last Ice Age ~19,000 years ago. During the major warming period after ~14,000 years ago, a genetic component related to present-day Near Easterners became widespread in Europe. These results document how population turnover and migration have been recurring themes of European prehistory

Modern humans arrived in Europe around 45,000 years ago and have   individuals from Europe[2–4]. Here we assemble and analyse ge

Artic

Cell

The genomic history of the Middle East

Graphical abstract

Physically Phased Genomes
Maternal ACTGAGACACTC
Paternal ACAGTGCCACTC
Accurate population history
Increase in variant frequency due to positive selection
Allele Frequency
Lactose tolerance
Mutation
Present
Past
Neanderthal Ancestry
Excess of Basal Eurasian and African ancestry
Arabia
Levant
Effective Population Size (Ne)
Levant
Arabia
Development of agriculture in the Levant
Desertification of Arabia
Years Ago
1e+03
1e+04
1e+05

Highlights

- Middle Easterners do not have ancestry from an early out-of-

Authors

Mohamed A. Almarri, Marc Haber, Reem A. Lootah, ..., Hilary C. Martin, Yali Xue, Chris Tyler-Smith

Correspondence

ma17@sanger.ac.uk (M.A.A.),
m.haber@bham.ac.uk (M.H.)

In brief

A high-coverage resource of physically phased genomes from eight Middle Eastern populations generated via linked-read sequencing provides insights into a genetically understudied region and enables more comprehensive study of population history and the detection of millions of variants common to the Middle East but outside short-read accessibility masks and not previously cataloged. It enhances our understanding of regional ancestry, the spread of languages, the effects of climate change on populations, and the evolutionary history of genetic variants.

But all the methods that you will learn during the course can be used for any organism, any population, any species

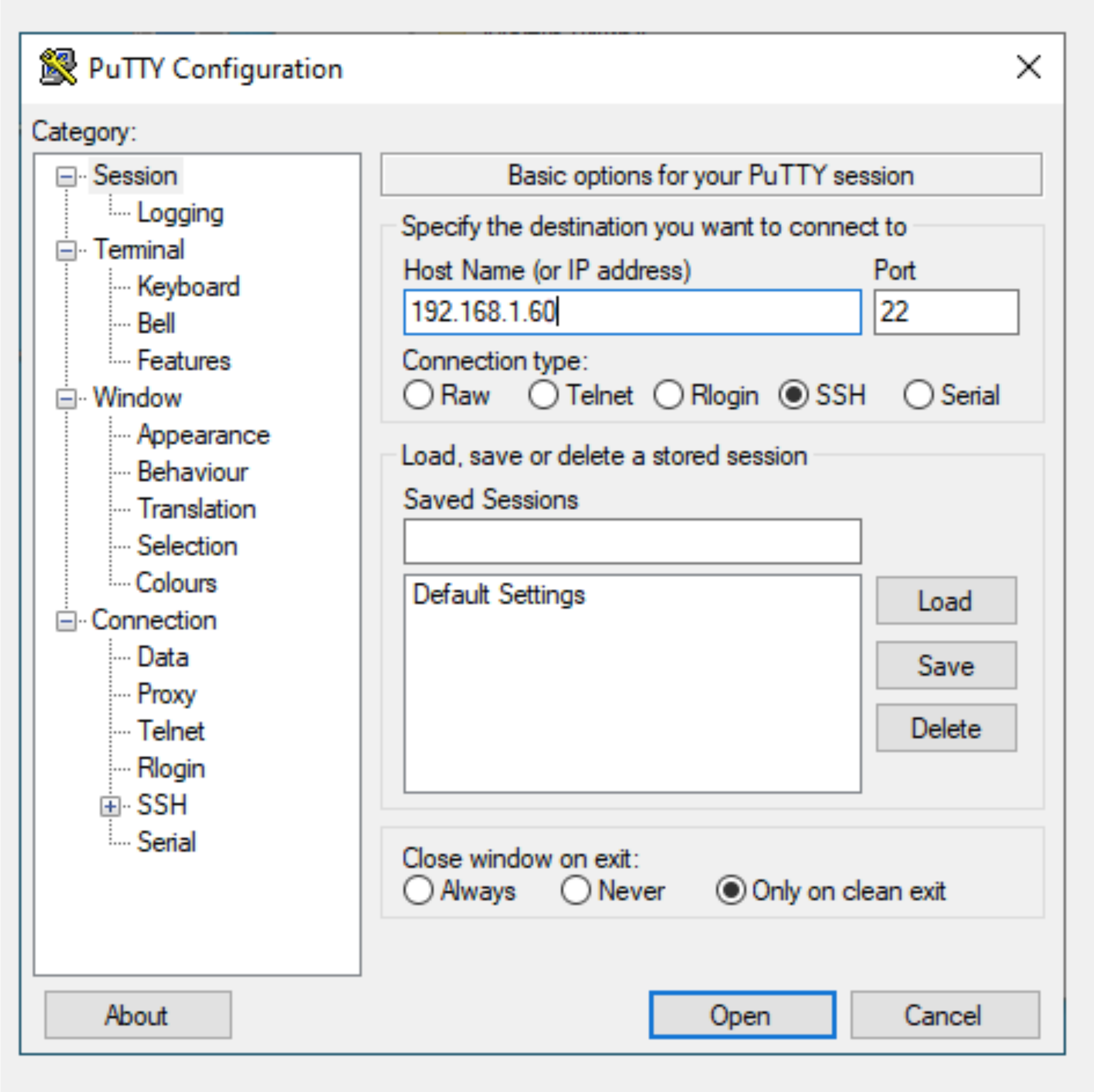
For the hands-on sessions:

Connect to the server:

Use terminal for ssh connection on Linux and Mac:

```
ssh yourusername@your.server.IP
```

PuTTy for ssh connection from Windows:



PuTTY Configuration

Basic options for your PuTTY session

Specify the destination you want to connect to

Host Name (or IP address): 192.168.1.60

Port: 22

Connection type: ○ Raw ○ Telnet ○ Rlogin ● SSH ○ Serial

Load, save or delete a stored session

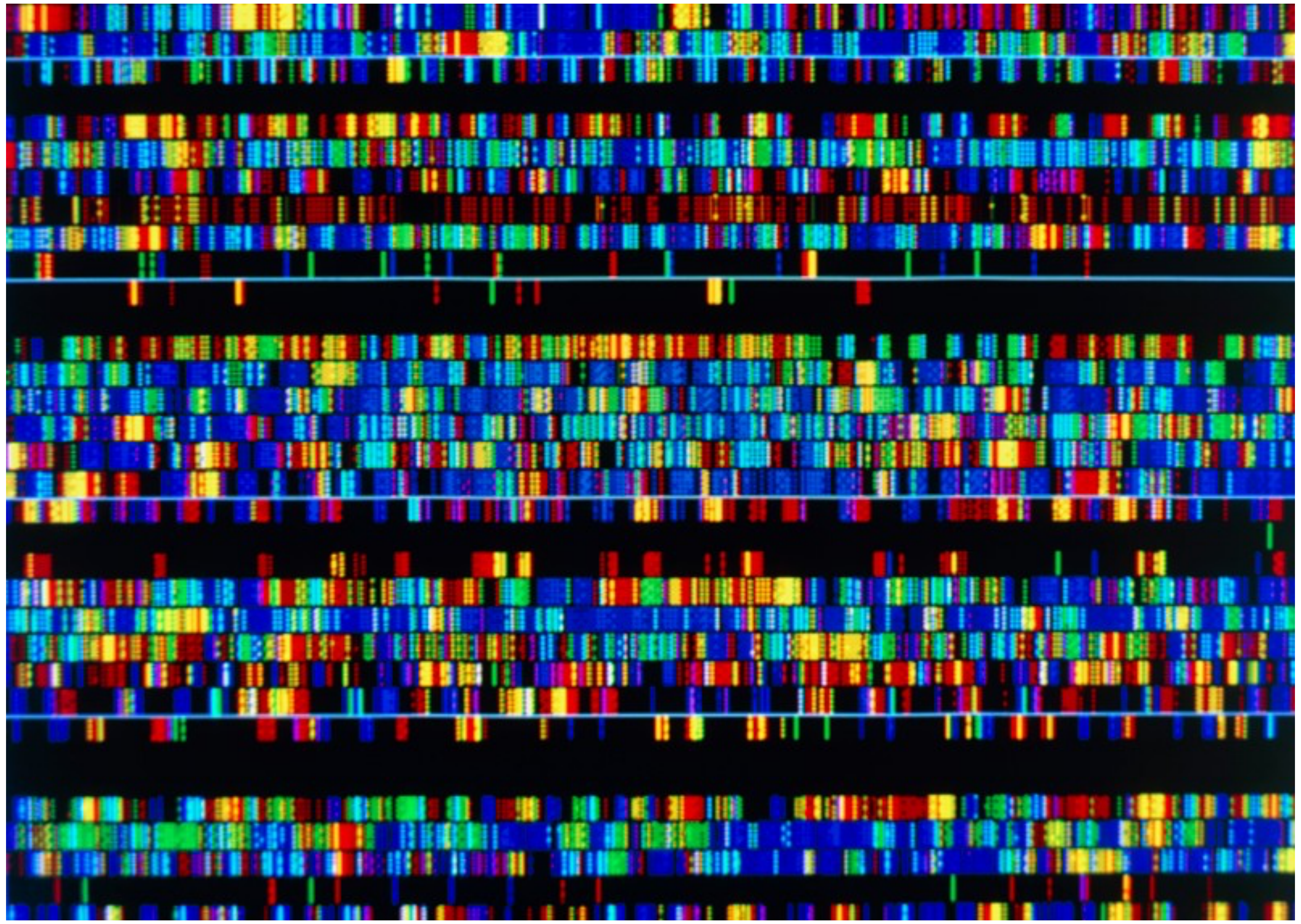
Saved Sessions: Default Settings

Close window on exit: ○ Always ○ Never ● Only on clean exit

Week-1 INTRODUCTION
Genome, sequencing and sequencing data - studying genetic variation

Aim: *Learning about the sequencing data, data formats, programs, softwares, tools to prepare genomic datasets for population genetics analyses*

Hands-on: *Examining the file formats, small edits on eigenstrat and plink files, conversion of file formats to each other*



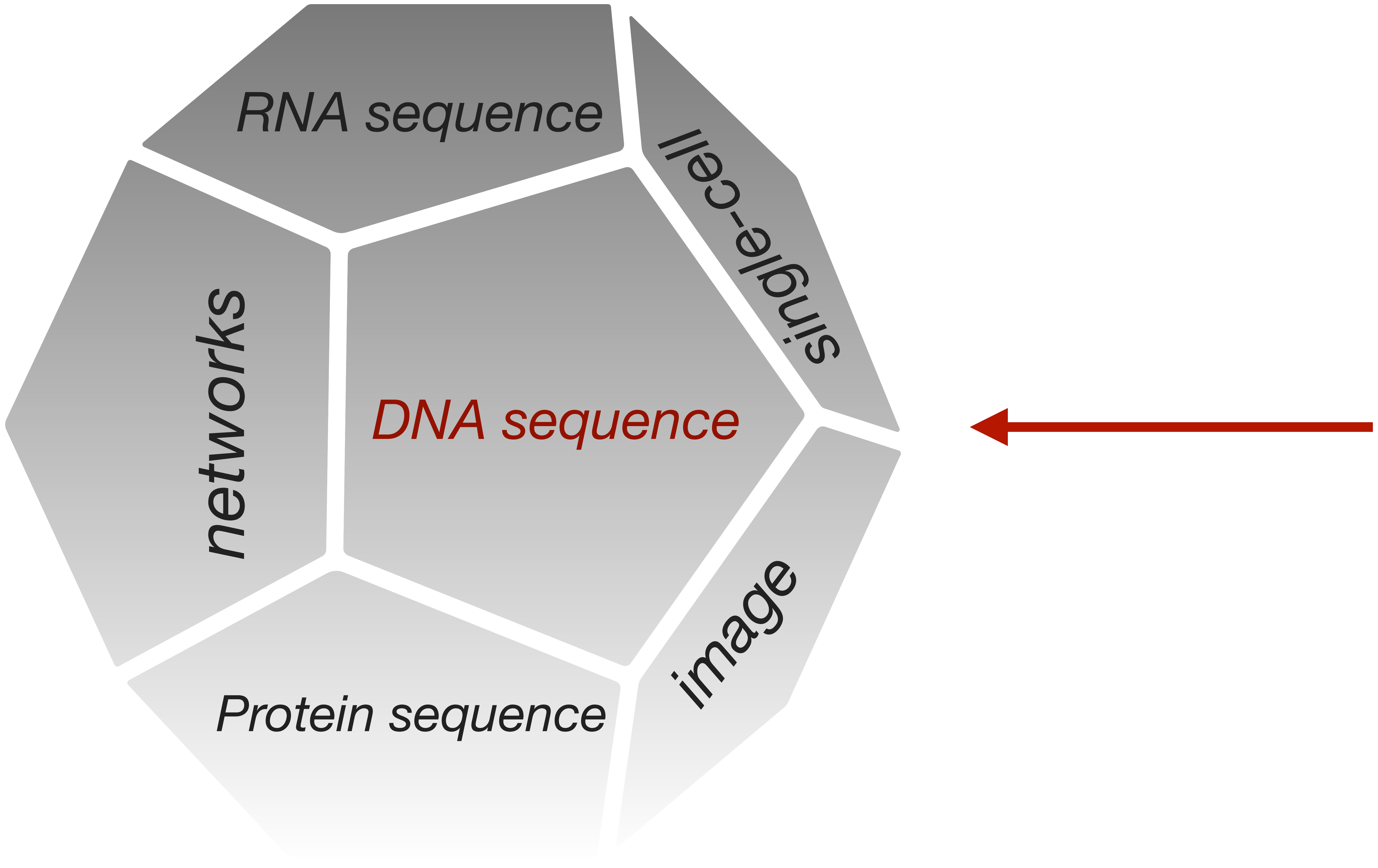
James King-Holmes/Science Photo Library, Nature, News 2021

Bioinformatics

Studying biological data

Population Genetics

Alleles in a population



RNA sequence

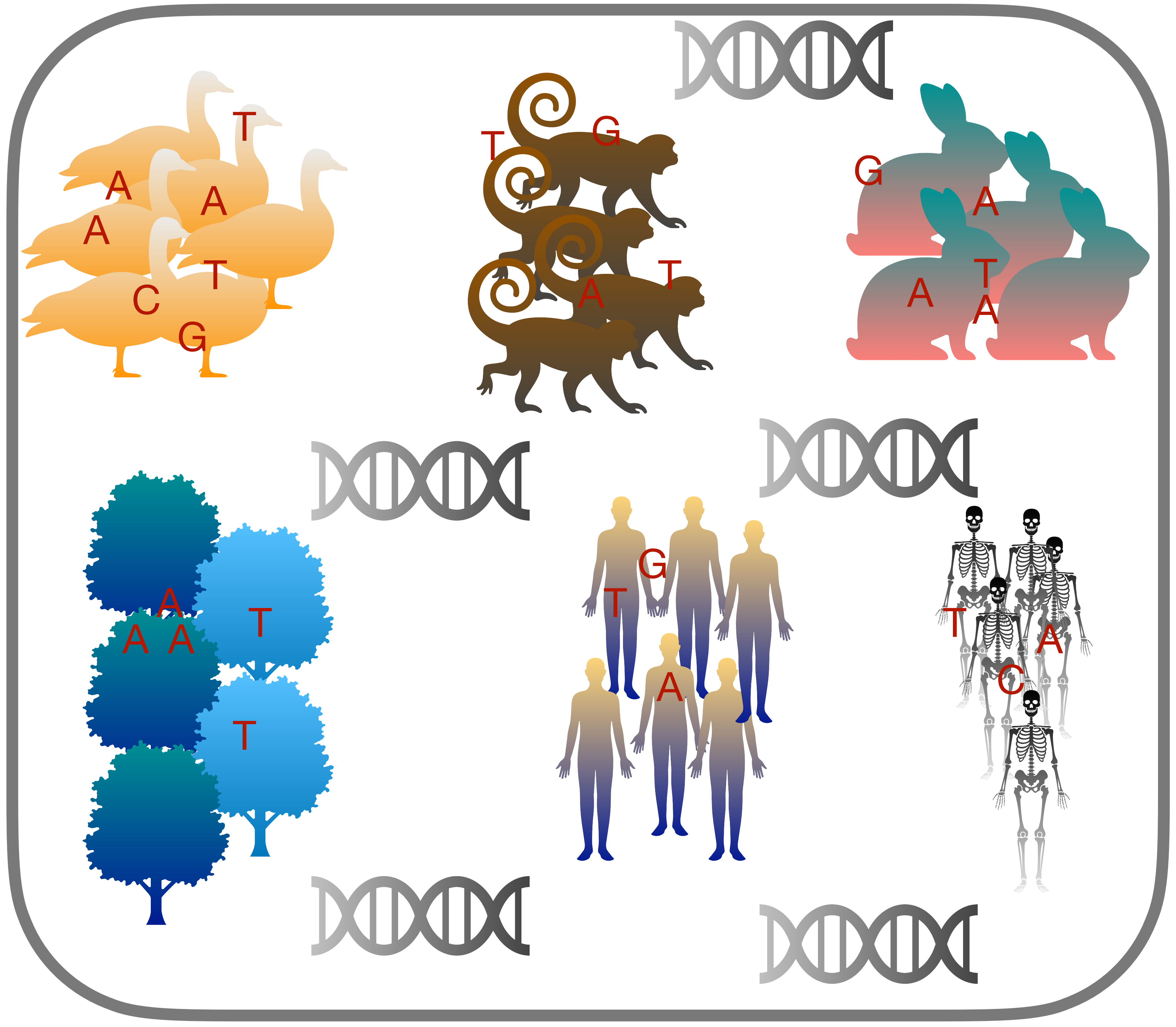
single-cell

networks

DNA sequence

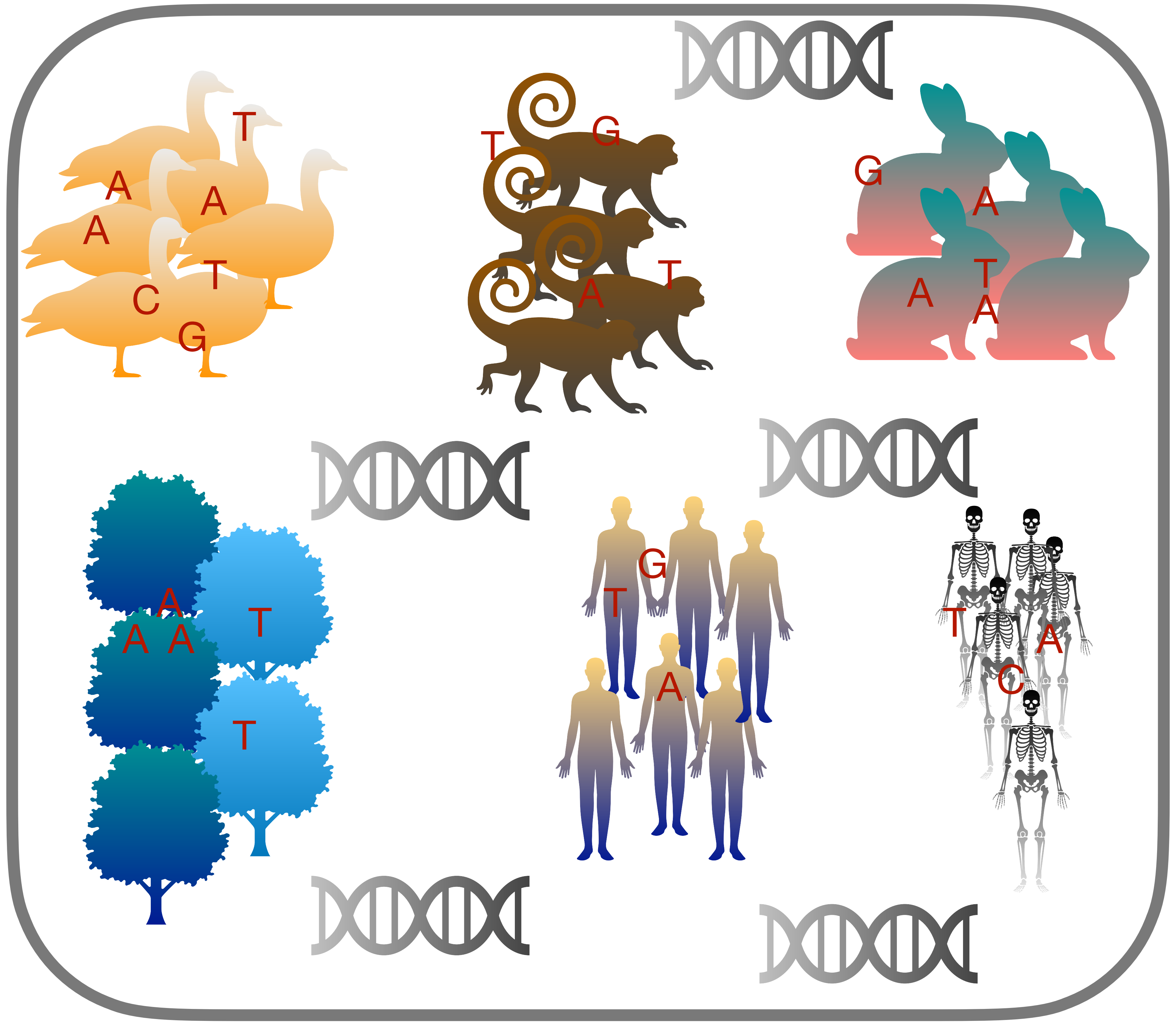
image

Protein sequence



Population Genetics

Alleles in a population

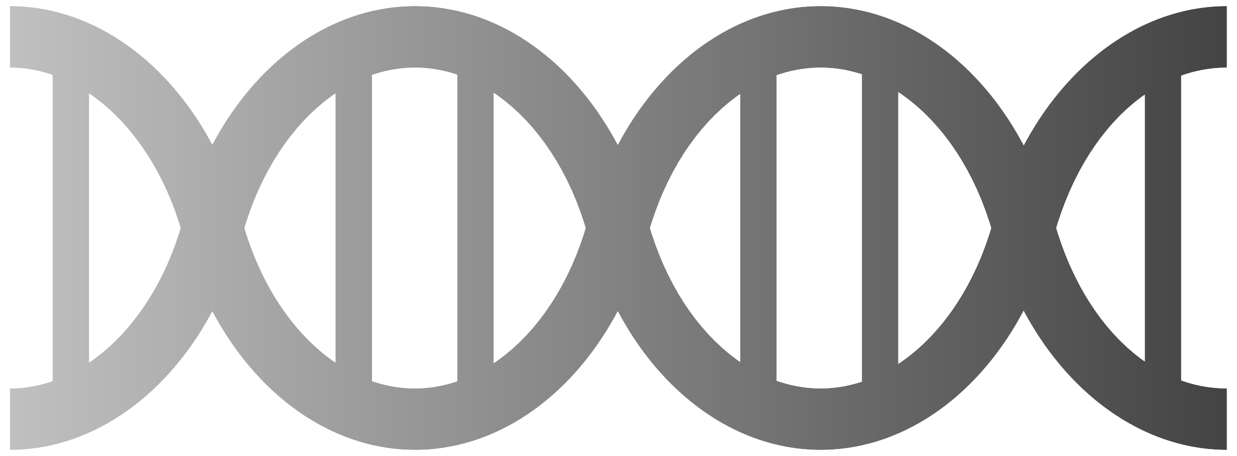


Alleles *-> genetic variants that are transmitted from parents to offsprings*

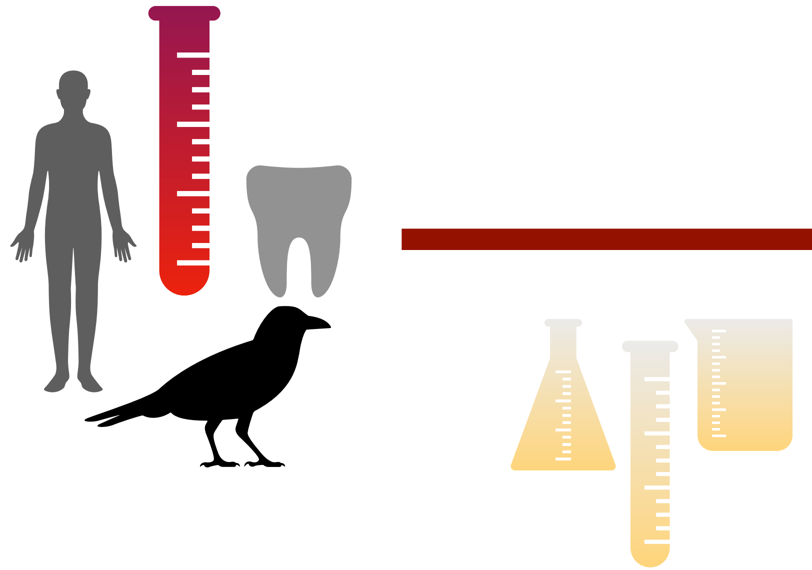
Types of genetic data

Single nucleotide polymorphism (SNP) C/T
Insertion/deletion CTATATCTCT -> CTAT—-TCT
Simple sequence repeats ATGCCACACATCG
Copy number variations

Genome sequencing



Finding the complete sequence of DNA:



AACTGTGCTGAGATGTCGTGTGCTAGAA

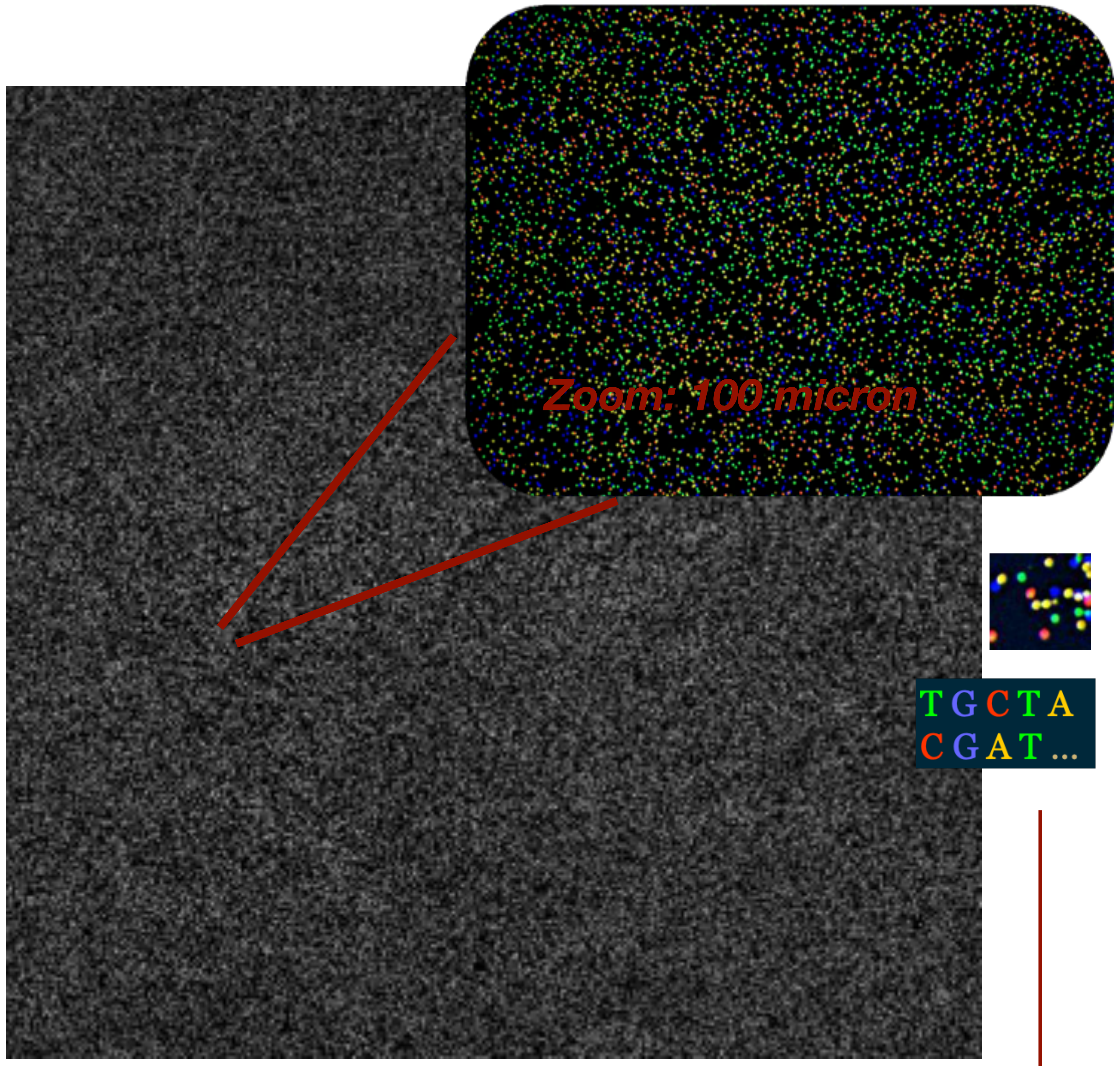
Analyse the ***data***

Bioinformatics

Genome Sequencing -> Next Generation Sequencing

Features of sequencing data

- Short sequence reads 35-150 bp (Illumina)
- Large amount of sequencing data (Up to gigabases per run)
- Large number of reads in each run (billions)
- GC bias
- High error rate compared to Sanger or compared to genotyping arrays



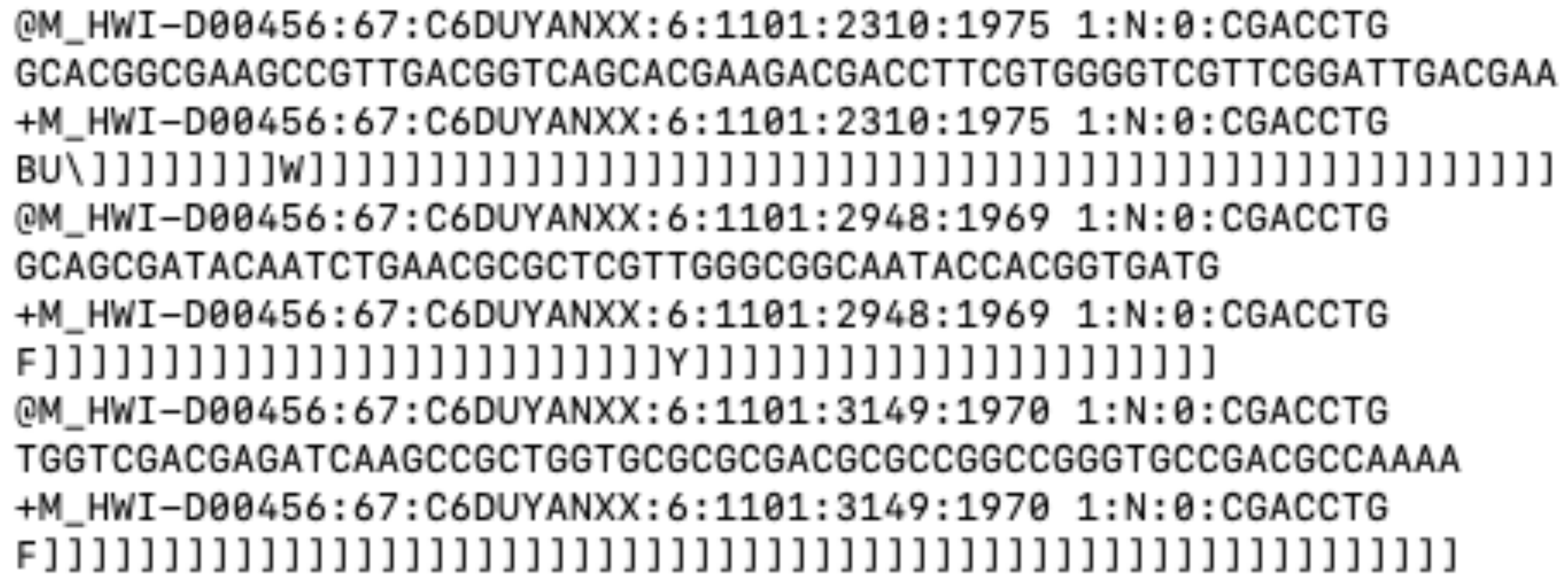
Raw NGS data: A snapshot from an Illumina image file

Base calling -> FASTQ file

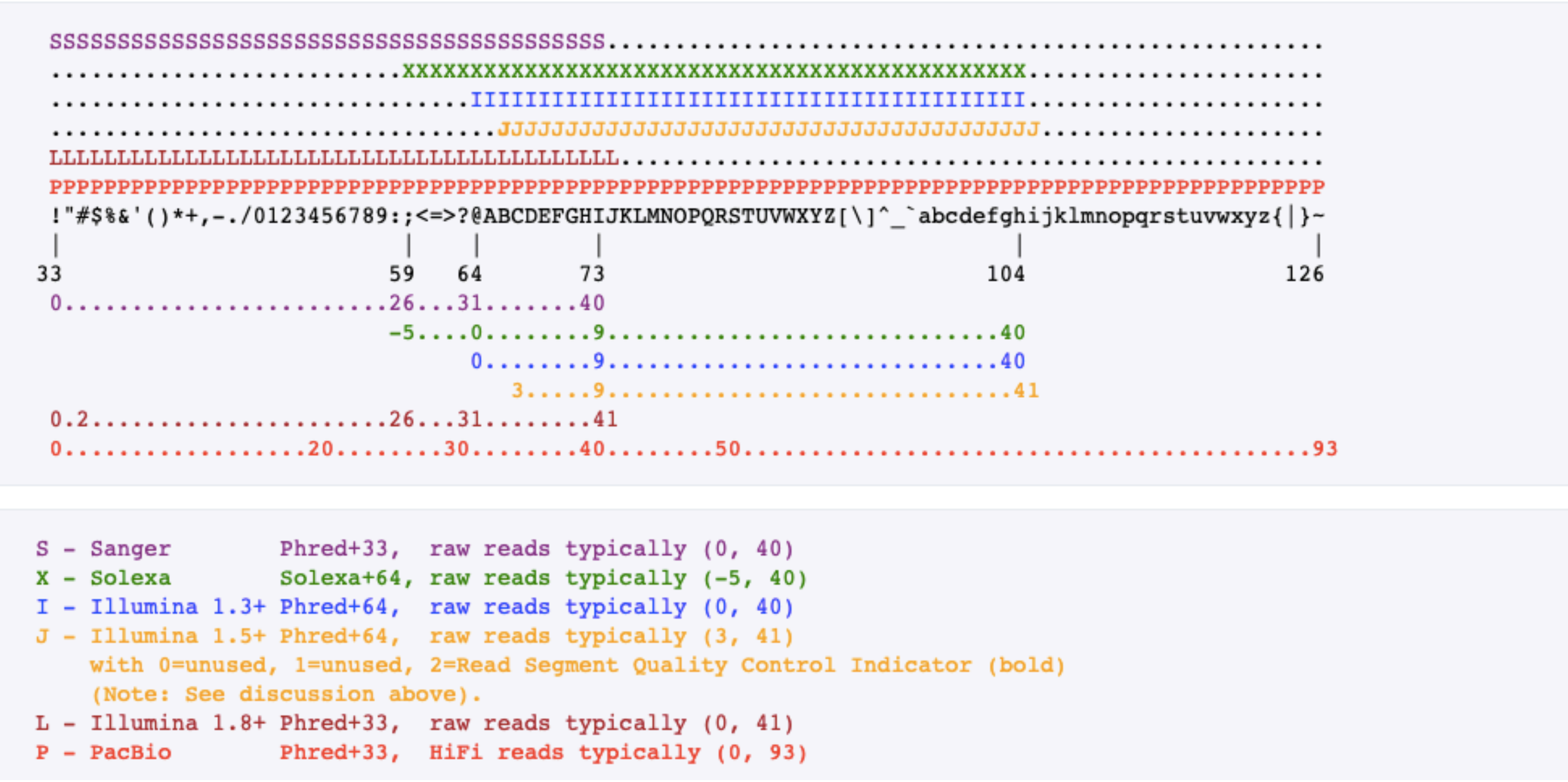
Genome Sequencing Data - (1) FASTQ

Raw unaligned read sequences - includes base qualities

- Sequence + base quality for each base of the sequence
- Subsets of ASCII printable characters
- https://en.wikipedia.org/wiki/FASTQ_format
- Line 1: begins with @ character, + sequence ID + description (optional)
- Line 2: sequence letters
- Line 3: + and same w/ Line 1
- Line 4: Quality values



```
@M_HWI-D00456:67:C6DUYANXX:6:1101:2310:1975 1:N:0:CGACCTG
GCACGGCGAAGCCGTTGACGGTCAGCACGAAGACGACCTTCGTGGGGTCGTTCGGATTGACGAA
+M_HWI-D00456:67:C6DUYANXX:6:1101:2310:1975 1:N:0:CGACCTG
BU\]]]]]]]]W]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]
@M_HWI-D00456:67:C6DUYANXX:6:1101:2948:1969 1:N:0:CGACCTG
GCAGCGATACAATCTGAACGCGCTCGTTGGGCGGCAATACCACGGTGATG
+M_HWI-D00456:67:C6DUYANXX:6:1101:2948:1969 1:N:0:CGACCTG
F]]]]]]]]]]]]]]]]]]]]]]]]]]Y]]]]]]]]]]]]]]]]]]]]]]
@M_HWI-D00456:67:C6DUYANXX:6:1101:3149:1970 1:N:0:CGACCTG
TGGTCGACGAGATCAAGCCGCTGGTGCGCGCGACGCGCCGGCCGGGTGCCGACGCCAAAA
+M_HWI-D00456:67:C6DUYANXX:6:1101:3149:1970 1:N:0:CGACCTG
F]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]
```



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0.2......................26...31........41
 0..................20........30........40........50.........................................93
```

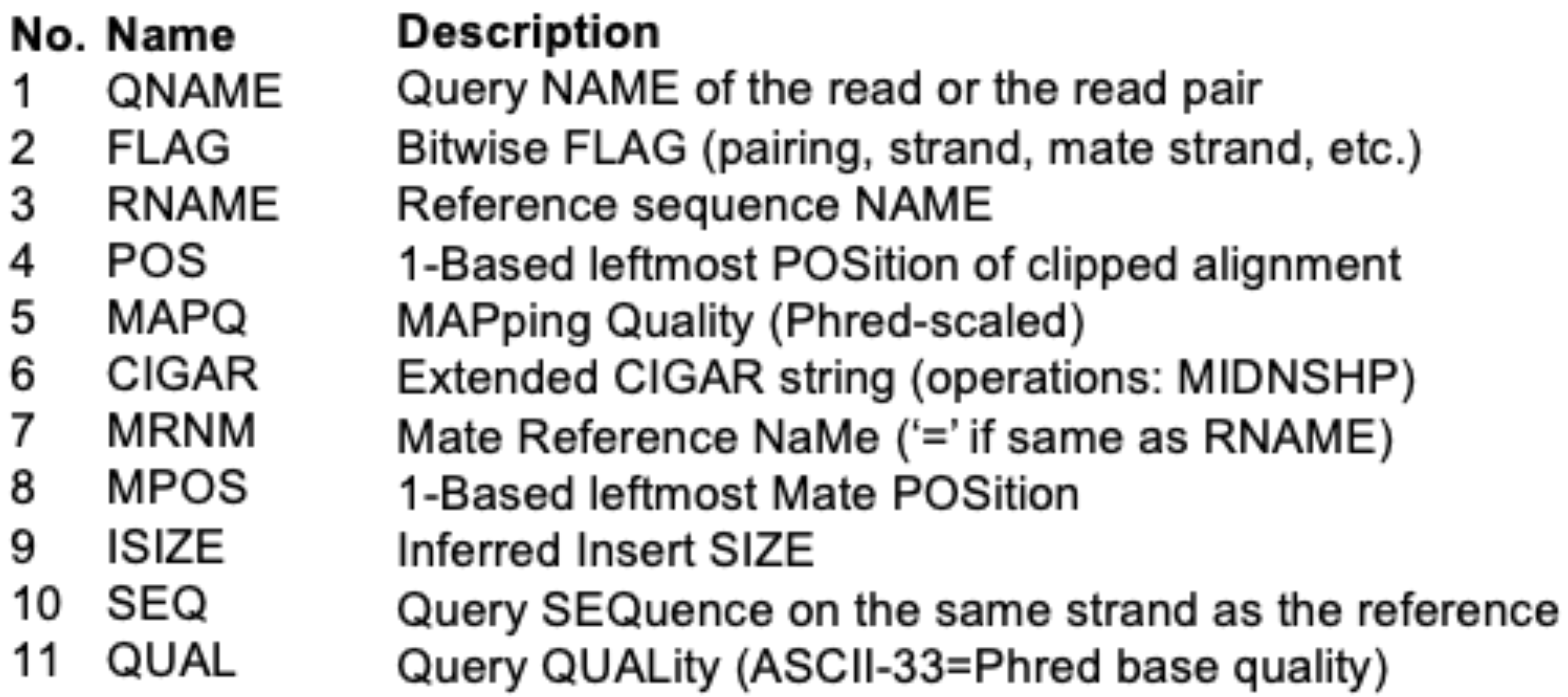
```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Genome Sequencing Data - (2) SAM/BAM

Sequence Alignment Map/ Binary Alignment/Map

- SAM -> Store read alignments to a reference genome
- BAM -> Binary format of SAM - for fast processing
- https://samtools.github.io/hts-specs/SAMv1.pdf
- Compact size
- Supported by variant calling softwares/tools
- Supports multiple sequencing technologies
- Reads can be grouped - lanes, libraries, samples - a more organised way of storing sequence data

Genome Sequencing Data - (2) SAM/BAM



No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

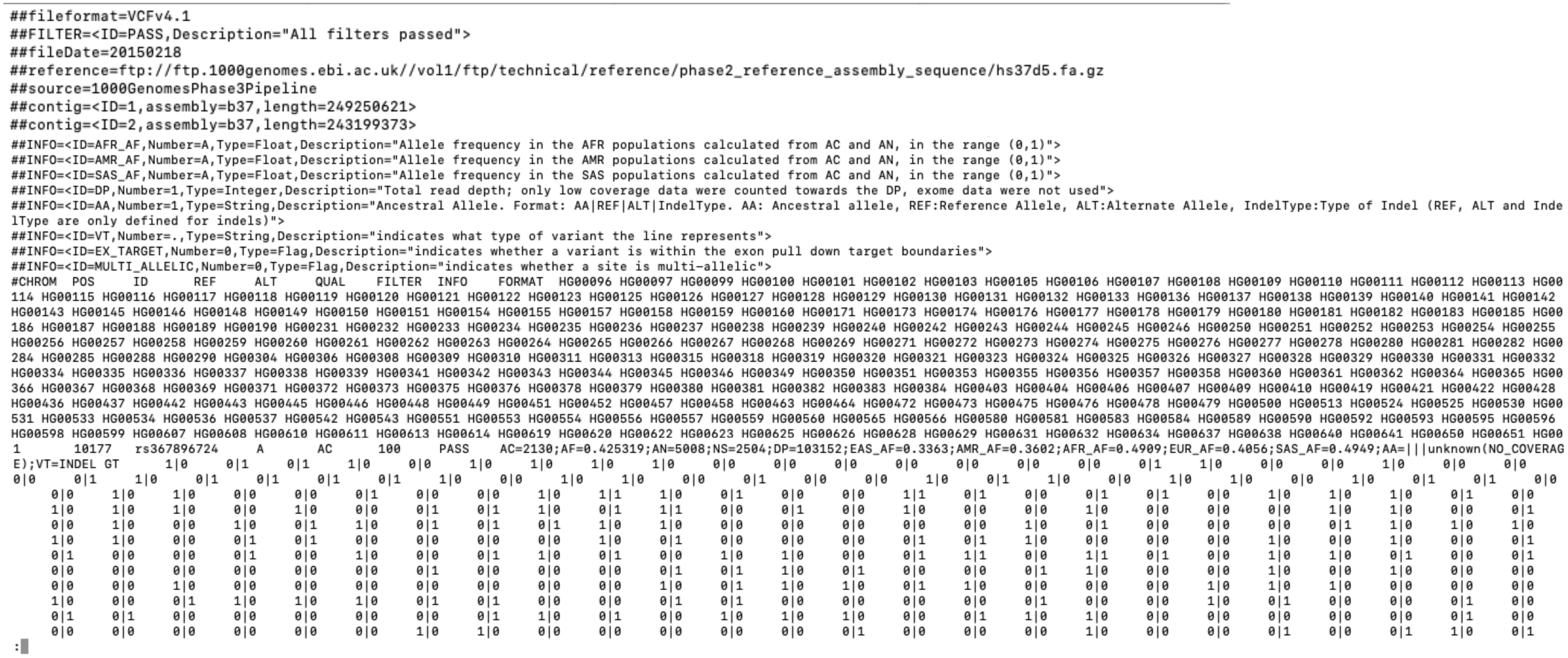
```
M_ST-E00198:315:HKWVGCCXY:4:1115:6695:8939      0       1       9995    0       79M     *
0       0       CCCTAATAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCT
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]] XT:A:R  XN:i:6
X0:i:2  X1:i:0  XM:i:4  XO:i:0  XG:i:0  XA:Z:12,-133841822,79M,4;       XP:i:3  NM:i:6  MD:Z:
0N0N0N0N0N0N73
```

Genotype Data - (3) Variant Call Format (VCF)

Standardized file format for storing the variant data

- SNPs, indels, structural variants - we use SNPs in the class
- Annotations for each variant
- https://samtools.github.io/hts-specs/VCFv4.2.pdf
- Compact size, many samples in the same file
- Meta data: filter status, variant access number (dbSNP)
- Flexible - user extended
- Structure: Header + Mandatory columns: CHR, POS, ID, REF, ALT, QUAL, FILTER, INFO

Genotype Data - (3) Variant Call Format (VCF)



```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##fileDate=20150218
##reference=ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz
##source=1000GenomesPhase3Pipeline
##contig=<ID=1,assembly=b37,length=249250621>
##contig=<ID=2,assembly=b37,length=243199373>
##INFO=<ID=AFR_AF,Number=A,Type=Float,Description="Allele frequency in the AFR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=AMR_AF,Number=A,Type=Float,Description="Allele frequency in the AMR populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=SAS_AF,Number=A,Type=Float,Description="Allele frequency in the SAS populations calculated from AC and AN, in the range (0,1)">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth; only low coverage data were counted towards the DP, exome data were not used">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele. Format: AA|REF|ALT|IndelType. AA: Ancestral allele, REF:Reference Allele, ALT:Alternate Allele, IndelType:Type of Indel (REF, ALT and Inde
lType are only defined for indels)">
##INFO=<ID=VT,Number=.,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=EX_TARGET,Number=0,Type=Flag,Description="indicates whether a variant is within the exon pull down target boundaries">
##INFO=<ID=MULTI_ALLELIC,Number=0,Type=Flag,Description="indicates whether a site is multi-allelic">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00105 HG00106 HG00107 HG00108 HG00109 HG00110 HG00111 HG00112 HG00113 HG00
114 HG00115 HG00116 HG00117 HG00118 HG00119 HG00120 HG00121 HG00122 HG00123 HG00125 HG00126 HG00127 HG00128 HG00129 HG00130 HG00131 HG00132 HG00133 HG00136 HG00137 HG00138 HG00139 HG00140 HG00141 HG00142
HG00143 HG00145 HG00146 HG00148 HG00149 HG00150 HG00151 HG00154 HG00155 HG00157 HG00158 HG00159 HG00160 HG00171 HG00173 HG00174 HG00176 HG00177 HG00178 HG00179 HG00180 HG00181 HG00182 HG00183 HG00185 HG00
186 HG00187 HG00188 HG00189 HG00190 HG00231 HG00232 HG00233 HG00234 HG00235 HG00236 HG00237 HG00238 HG00239 HG00240 HG00242 HG00243 HG00244 HG00245 HG00246 HG00250 HG00251 HG00252 HG00253 HG00254 HG00255
HG00256 HG00257 HG00258 HG00259 HG00260 HG00261 HG00262 HG00263 HG00264 HG00265 HG00266 HG00267 HG00268 HG00269 HG00271 HG00272 HG00273 HG00274 HG00275 HG00276 HG00277 HG00278 HG00280 HG00281 HG00282 HG00
284 HG00285 HG00288 HG00290 HG00304 HG00306 HG00308 HG00309 HG00310 HG00311 HG00313 HG00315 HG00318 HG00319 HG00320 HG00321 HG00323 HG00324 HG00325 HG00326 HG00327 HG00328 HG00329 HG00330 HG00331 HG00332
HG00334 HG00335 HG00336 HG00337 HG00338 HG00339 HG00341 HG00342 HG00343 HG00344 HG00345 HG00346 HG00349 HG00350 HG00351 HG00353 HG00355 HG00356 HG00357 HG00358 HG00360 HG00361 HG00362 HG00364 HG00365 HG00
366 HG00367 HG00368 HG00369 HG00371 HG00372 HG00373 HG00375 HG00376 HG00378 HG00379 HG00380 HG00381 HG00382 HG00383 HG00384 HG00403 HG00404 HG00406 HG00407 HG00409 HG00410 HG00419 HG00421 HG00422 HG00428
HG00436 HG00437 HG00442 HG00443 HG00445 HG00446 HG00448 HG00449 HG00451 HG00452 HG00457 HG00458 HG00463 HG00464 HG00472 HG00473 HG00475 HG00476 HG00478 HG00479 HG00500 HG00513 HG00524 HG00525 HG00530 HG00
531 HG00533 HG00534 HG00536 HG00537 HG00542 HG00543 HG00551 HG00553 HG00554 HG00556 HG00557 HG00559 HG00560 HG00565 HG00566 HG00580 HG00581 HG00583 HG00584 HG00589 HG00590 HG00592 HG00593 HG00595 HG00596
HG00598 HG00599 HG00607 HG00608 HG00610 HG00611 HG00613 HG00614 HG00619 HG00620 HG00622 HG00623 HG00625 HG00626 HG00628 HG00629 HG00631 HG00632 HG00634 HG00637 HG00638 HG00640 HG00641 HG00650 HG00651 HG00
1       10177   rs367896724     A       AC      100     PASS    AC=2130;AF=0.425319;AN=5008;NS=2504;DP=103152;EAS_AF=0.3363;AMR_AF=0.3602;AFR_AF=0.4909;EUR_AF=0.4056;SAS_AF=0.4949;AA=|||unknown(NO_COVERAG
E);VT=INDEL GT      1|0     0|1     0|1     1|0     0|0     1|0     1|0     1|0     1|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|1     1|0     0|0     0|0     1|0     0|0     0|0
0|0     0|1     1|0     0|1     0|1     0|1     0|1     1|0     0|0     1|0     1|0     0|0     0|1     0|0     0|0     1|0     0|1     1|0     0|0     1|0     1|0     0|0     1|0     0|1     0|1     0|0
      0|0     1|0     1|0     0|0     0|0     0|1     0|0     0|0     1|0     1|1     1|0     0|1     0|0     0|0     1|1     0|1     0|0     0|1     0|1     0|0     1|0     1|0     1|0     0|1     0|0
      1|0     1|0     1|0     0|0     1|0     0|0     0|1     0|1     1|0     0|1     1|1     0|0     0|1     0|0     1|0     0|0     0|0     1|0     0|0     0|0     0|0     1|0     1|0     0|0     0|1
      0|0     1|0     0|0     1|0     0|1     1|0     0|1     0|1     0|1     1|0     1|0     0|0     0|0     0|0     0|0     0|0     1|0     0|1     0|0     0|0     0|0     0|1     1|0     1|0     1|0
      1|0     1|0     0|0     0|1     0|1     0|0     0|0     0|0     0|0     1|0     0|1     0|0     0|0     0|0     0|1     0|1     1|0     0|0     0|0     0|0     1|0     0|0     1|0     0|0     0|1
      0|1     0|0     0|0     0|1     0|0     1|0     0|0     0|1     1|0     0|1     0|0     1|0     1|0     0|0     0|1     1|1     0|0     1|1     0|1     0|0     1|0     1|0     0|1     0|0     0|1
      0|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0     0|0     0|0     0|1     0|1     1|0     0|1     0|0     0|0     0|1     1|0     0|0     0|0     1|0     0|0     1|0     0|0     0|0
      0|0     0|0     1|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     1|0     0|1     1|0     1|0     0|1     1|0     0|0     0|0     0|0     1|0     1|0     1|0     0|0     0|0     0|0
      1|0     0|0     0|1     1|0     1|0     1|0     0|1     0|1     0|0     0|0     0|1     0|1     0|0     0|0     0|0     0|0     0|1     0|0     0|0     1|0     0|1     0|0     0|0     0|1     0|0
      0|1     0|1     0|0     0|0     0|0     0|0     0|0     0|1     1|0     0|1     0|0     1|0     1|0     1|0     0|0     0|1     1|0     1|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0
      0|0     0|0     0|0     0|0     0|0     0|0     1|0     1|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0     0|0     0|0     1|0     0|0     0|0     0|1     0|0     0|1     1|0     0|1
:
```

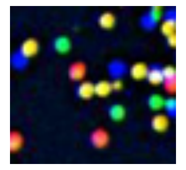
Genotype Data - Other file formats - mostly used in popgen analysis

Sequence Alignment Map/ Binary Alignment/Map

- mpileup
- PLINK -> ped, map, pedant
- EIGENSTRAT -> geno, snp, ind
- Plink & Eigenstrat -> file sets including three different files
- Can store population information
- Compatible with popgun tools/programs
- Can be easily converted to each other.

Workflow: How do we produce these files?

Softwares/tools/programs - what do we need?



Base calling -> FASTQ file

Map to the reference genome : BWA http://bio-bwa.sourceforge.net

BAM file

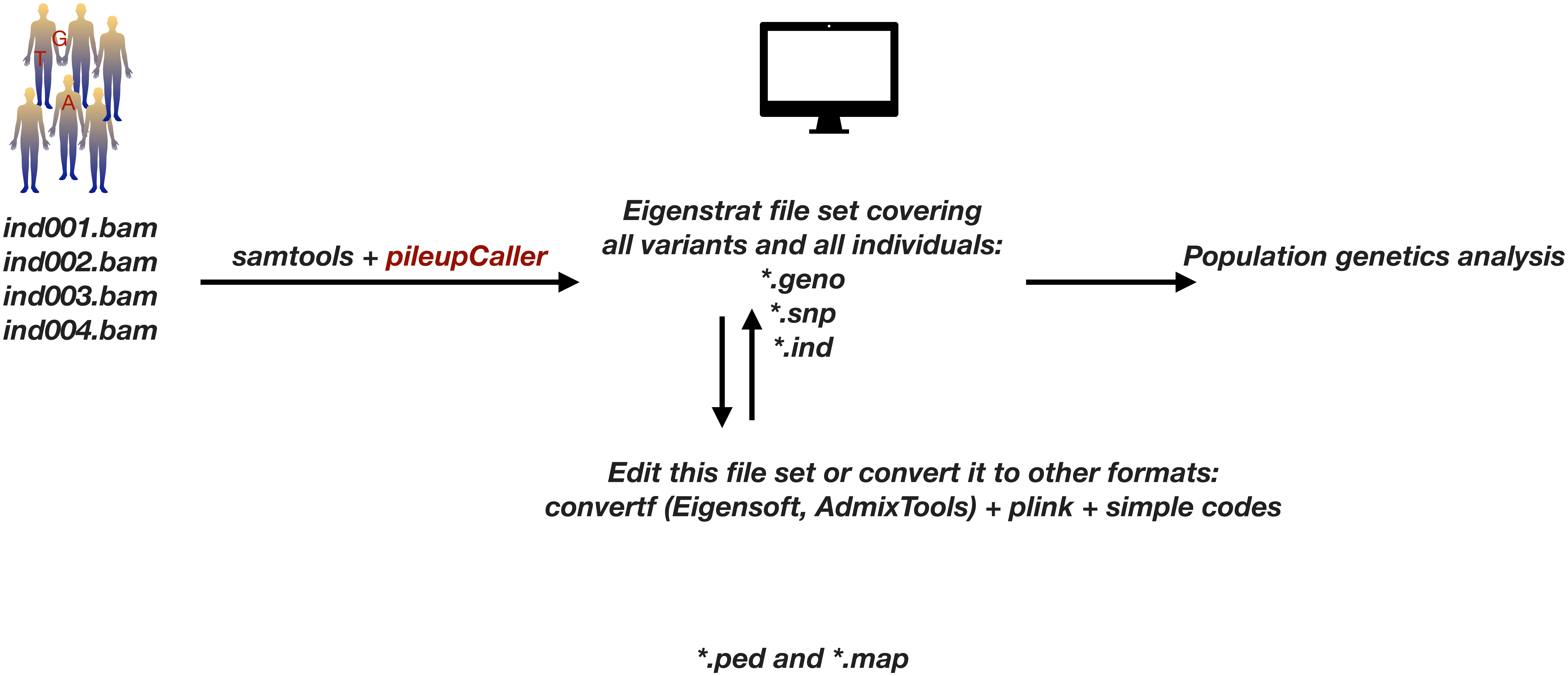
Filter/Discover the variants: samtools, GATK, pileupCaller, plink

VCF file, mpileup, eigenstrat, plink

File format conversion: AdmixTools -> convertf

Datasets: Preparing, converting, editing

Softwares/tools/programs - what do we need?



ind001.bam
ind002.bam
ind003.bam
ind004.bam

samtools + pileupCaller

Eigenstrat file set covering all variants and all individuals:
****.geno***
****.snp***
****.ind***

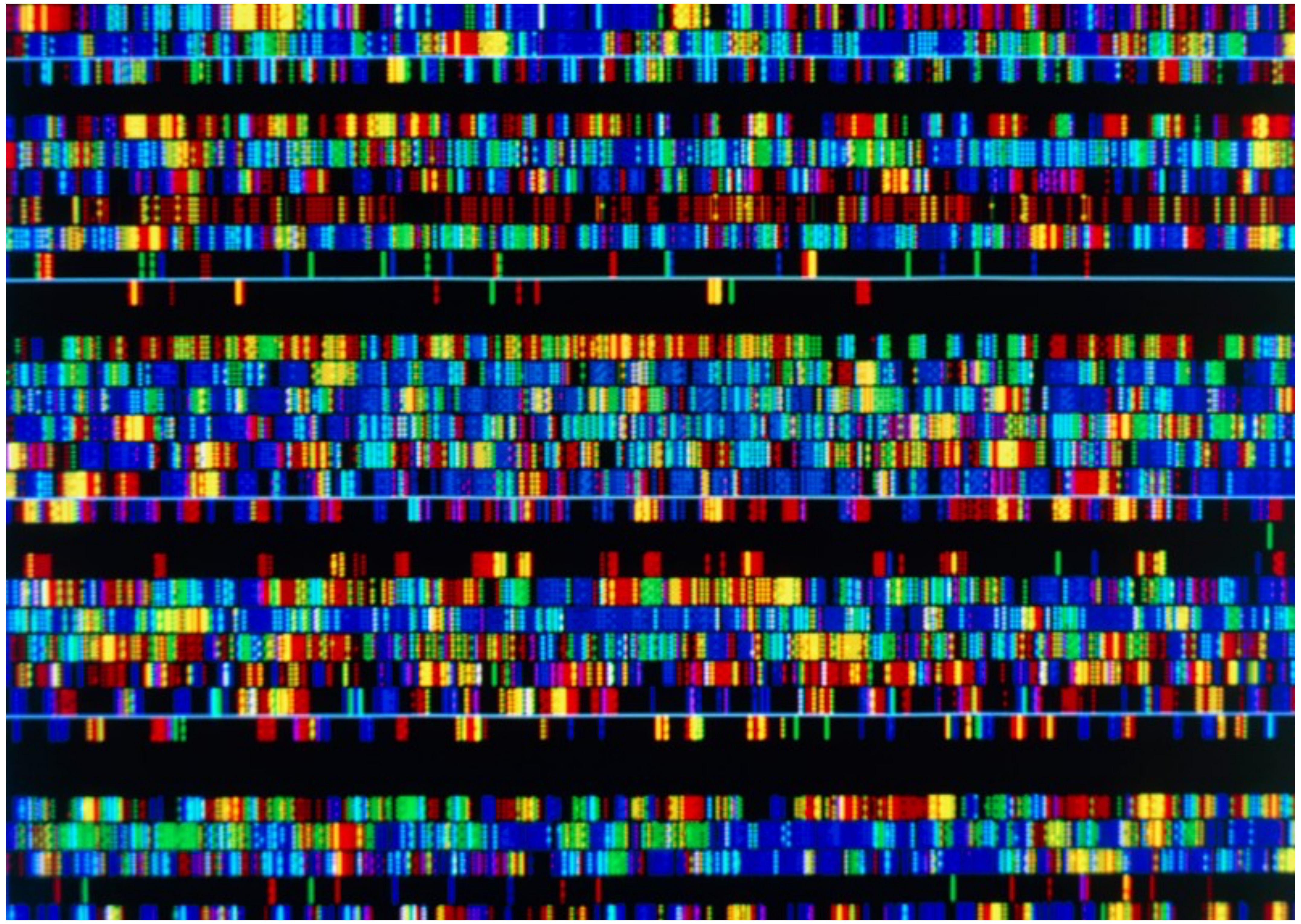
Population genetics analysis

Edit this file set or convert it to other formats:
convertf (Eigensoft, AdmixTools) + plink + simple codes

****.ped and *.map***

Week-1
Hands-on

Aim: *Becoming familiar with file formats, file format conversion, writing simple scripts to play with files*



James King-Holmes/Science Photo Library, Nature, News 2021

FASTQ File

zcat Sample1_Example1.fastq.gz | less

```
@M_HWI-D00456:67:C6DUYANXX:6:1101:1411:1958 1:N:0:CGACCTG
NCACGCCGCTTTCCTGCGGACTGTGGCGCGCCGTCACGATCAGCGCGCCGGCGCCGAAGGCGACCGCC
GAACGCAGGATCGCGCCGACATTGTGTGGATCGGTCACCTGGTCGAGCACGACCACCAGCGGCGCGTC
+M_HWI-D00456:67:C6DUYANXX:6:1101:1411:1958 1:N:0:CGACCTG
#<<BBFFFFF]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]
]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]]FFFFFBBBBB
```

BAM file

samtools view -h File1.bam | less -S

samtools view -h File2.bam | less -S

VCF file

```
less ALL.chr21.phase3_shapeit2_mvncall_integrated_v5a.
20130502.genotypes.vcf
```

EIGENSTRAT & PLINK Files

```
less -S data.ped
less data.pedind
less data.map

less data.snp
less data.ind
less data.geno
```

Convert datasets PED -> EIGENSTRAT -> PED and more…

Use: convertf [AdmixTools]

For more formats: https://github.com/chrchang/eigensoft/blob/master/CONVERTF/README

We need: A parameter file

Example 1:

```
genotypename: data.geno
snpname:      data.snp
indivname:    data.ind
outputformat: PED
genooutfilename:   data.ped
snpoutfilename:    data.map
indoutfilename:    data.pedind
outputgroup: YES
familynames: NO
hashcheck: NO
allowdups:  YES
pordercheck: NO
```

Example 2:

```
genotypename: data.ped
snpname:      data.map
indivname:    data.ped
outputformat: EIGENSTRAT
genooutfilename:   data.geno
snpoutfilename:    data.snp
indoutfilename:    data.ind
outputgroup: YES
familynames: NO
hashcheck: NO
allowdups:  YES
pordercheck: NO
```