



Centre for
Tropical Livestock
Genetics and Health

Genomic data management

Getinet M. Tarekegn, PhD



ROSLIN

SRUC

ILRI

CGIAR

Who am I?



Centre for
Tropical Livestock
Genetics and Health

Academic background:

- BSc in Animal SciencesHaramaya University
- MSc in Animal Genetics and BreedingHaramaya University
- BSc in Computer SciencesBahir Dar University
- Postgraduate studies in Higher Education.....Leeds Met. University, UK
- Visiting researcher in Small Ruminant Genomics ...Inner Mongolian Agri.University, China
- PhD in Applied Genetics (Livestock Genomics).... AAU, 2016
- Post-doc/Visiting researcher in African Goat Genomics ... BecA-ILRI, 2016/2017
- Post-doc in Dairy cattle genetics and genomics..... SLU, Sweden, 2017-2020
- Geneticist and Bioinformatician, Scotland's Rural College (SRUC)... 2022 and onwards

Affiliations:

- ✓ Adjunct Prof of Livestock Genomics and Bioinformatics, Addis Ababa University, Ethiopia
- ✓ Geneticist and Bioinformatician, Scotland's Rural College (SRUC),
Roslin Institute Building, University of Edinburgh, UK



Outline

- SNP genotypes
- Quality control
- Converting the snps data to different formats
- Plink practicals

Learning objectives:

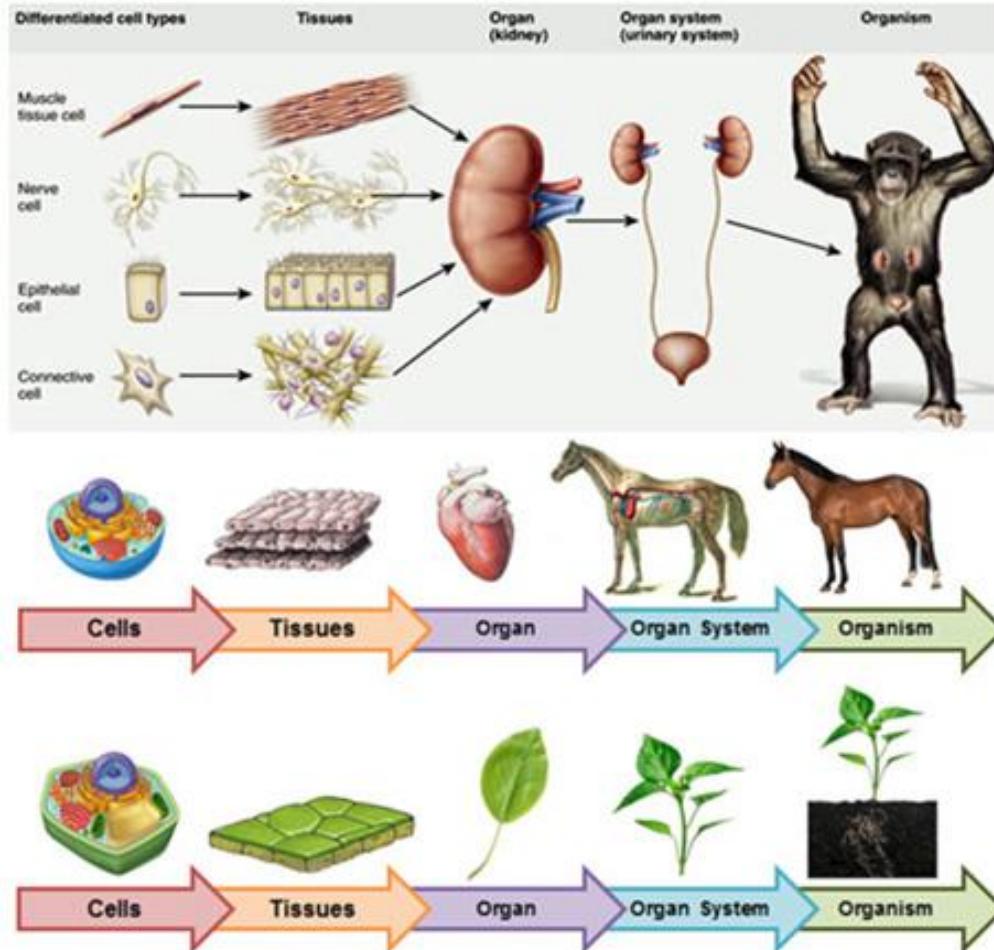
- Understand how to filter genotype data and converting to different formats

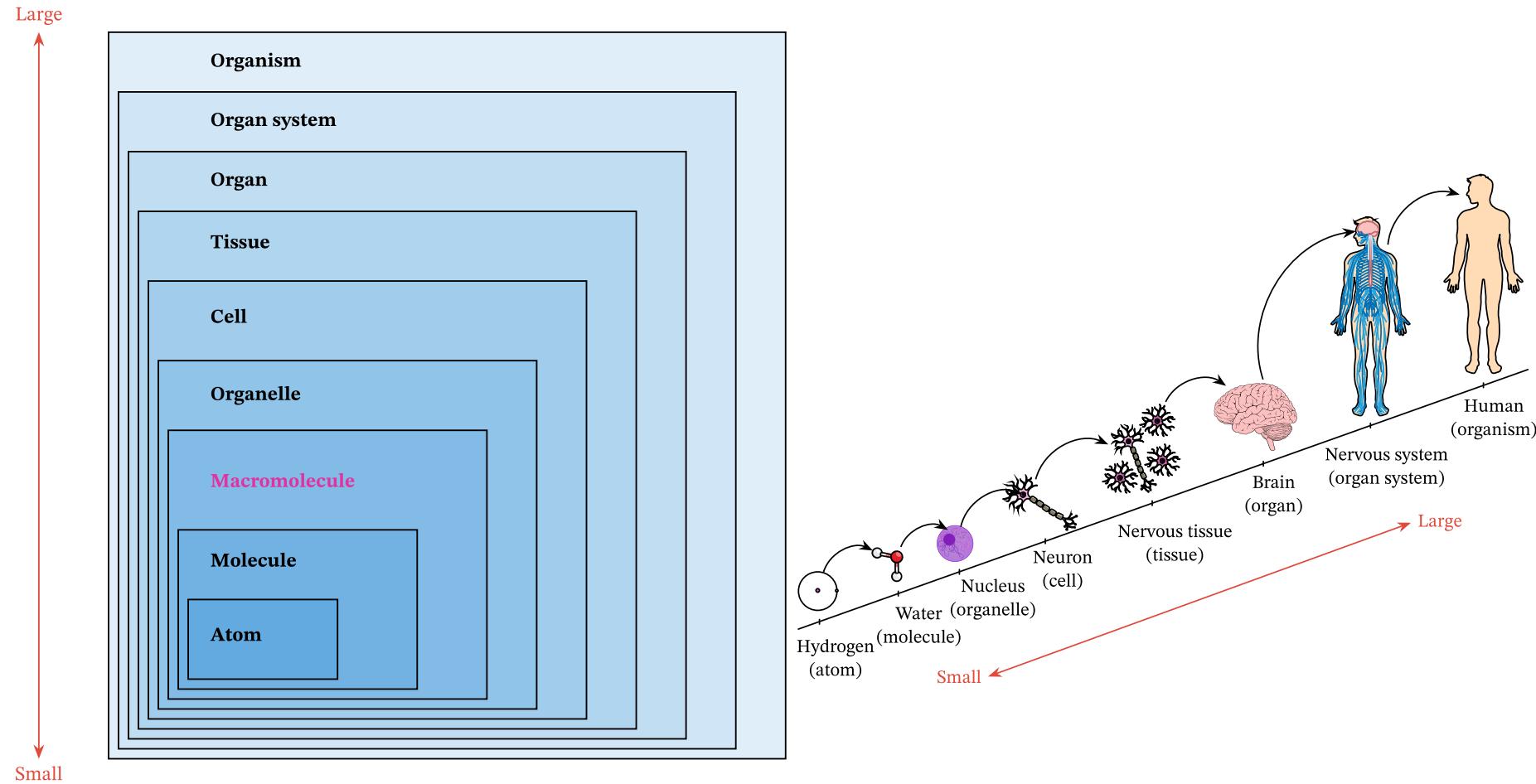
Learning outcome:

- All trainees are capable of handling genotype data



Biological levels of Organization of living things



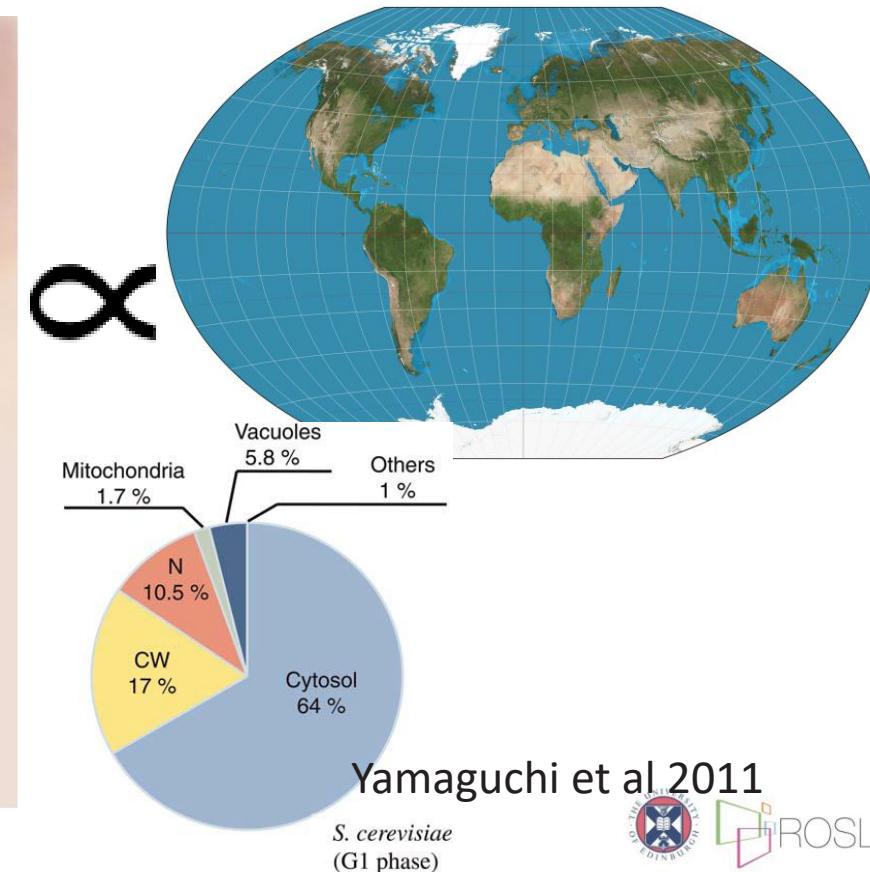
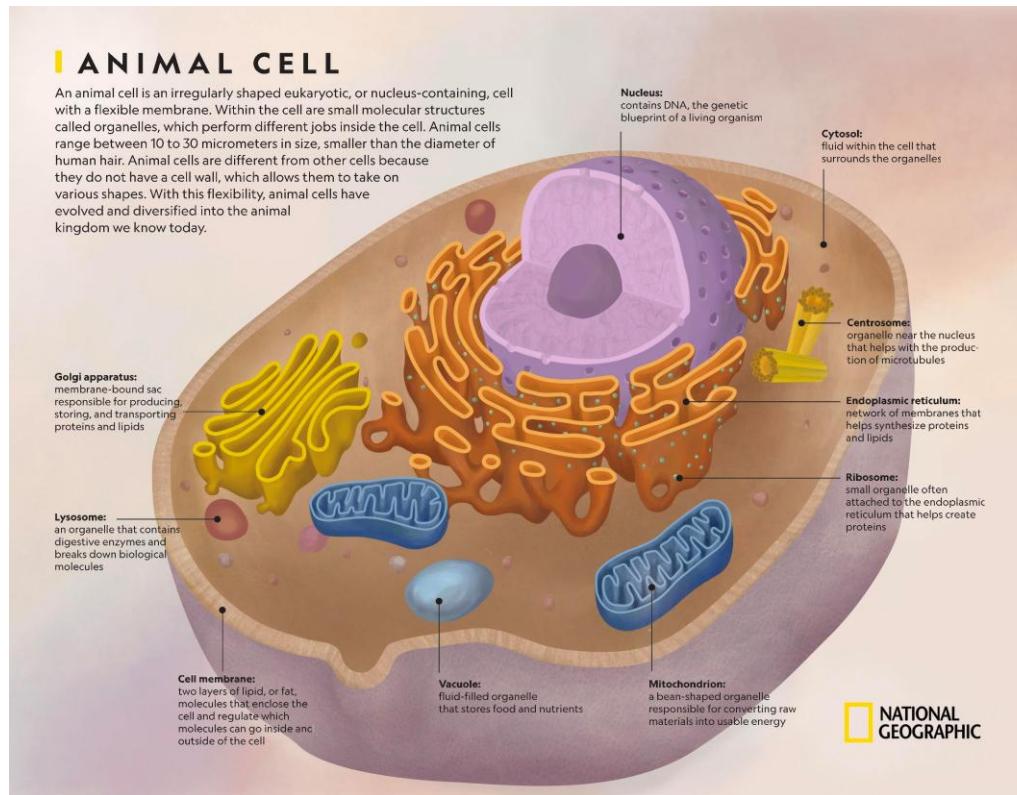


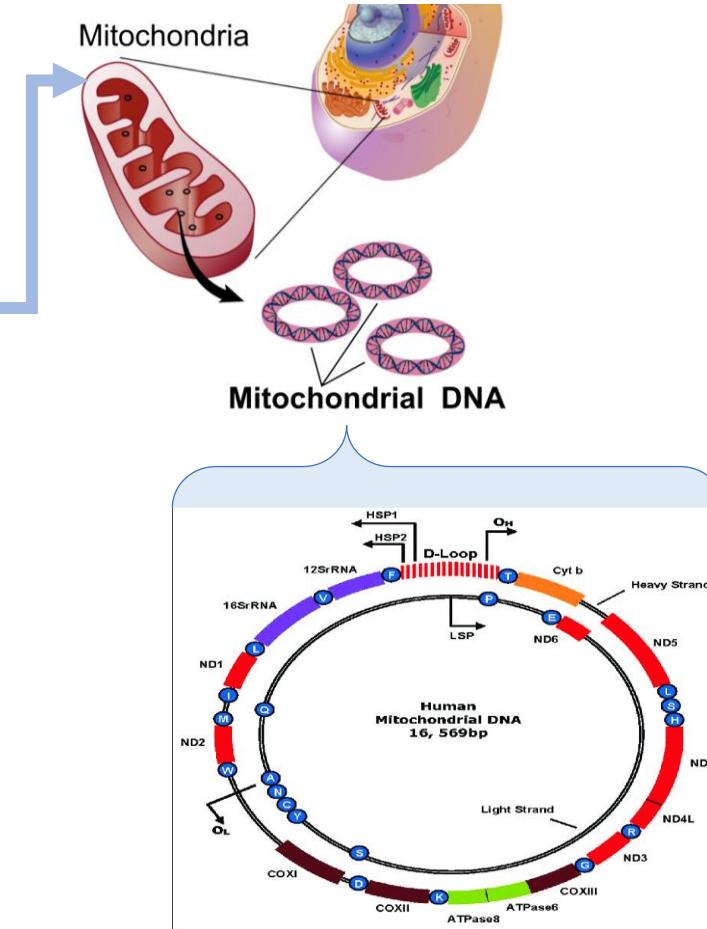
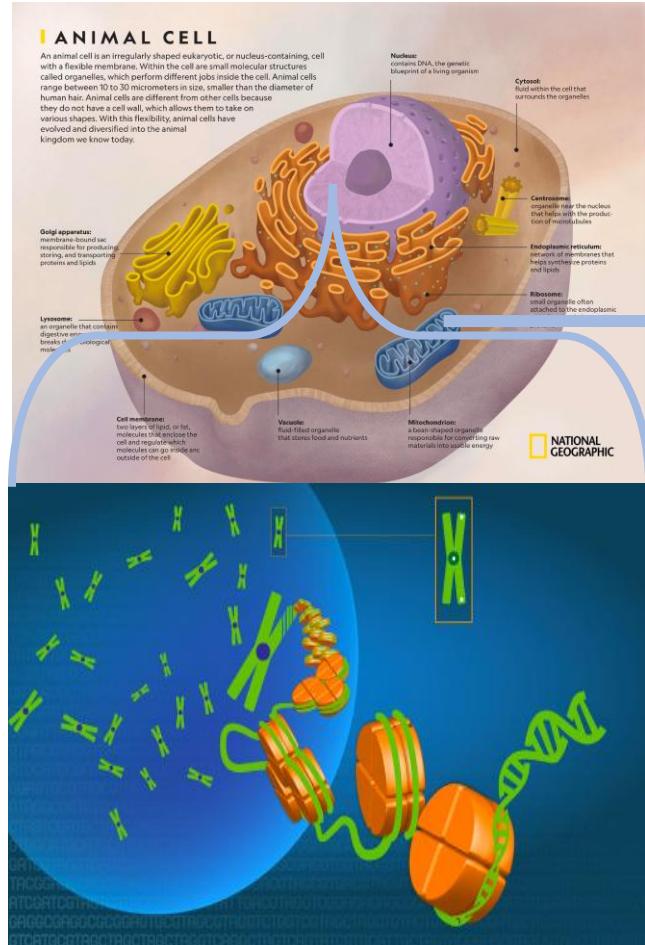
A diagram showing the biological levels of organization from an atom to a multicellular organism <https://www.nagwa.com/en/explainers/430187521519/>



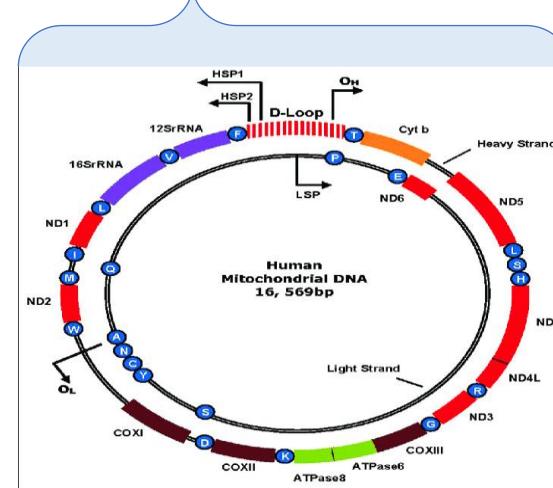
Cell is the smallest unit of life.

- can divide, multiply, grow and respond to stimuli from the environment; but is it as smallest in size as we expect?





- 5 cm long (about 2 inches) each chr, and all 46 chrs be about 2 m





Interesting Facts about DNA

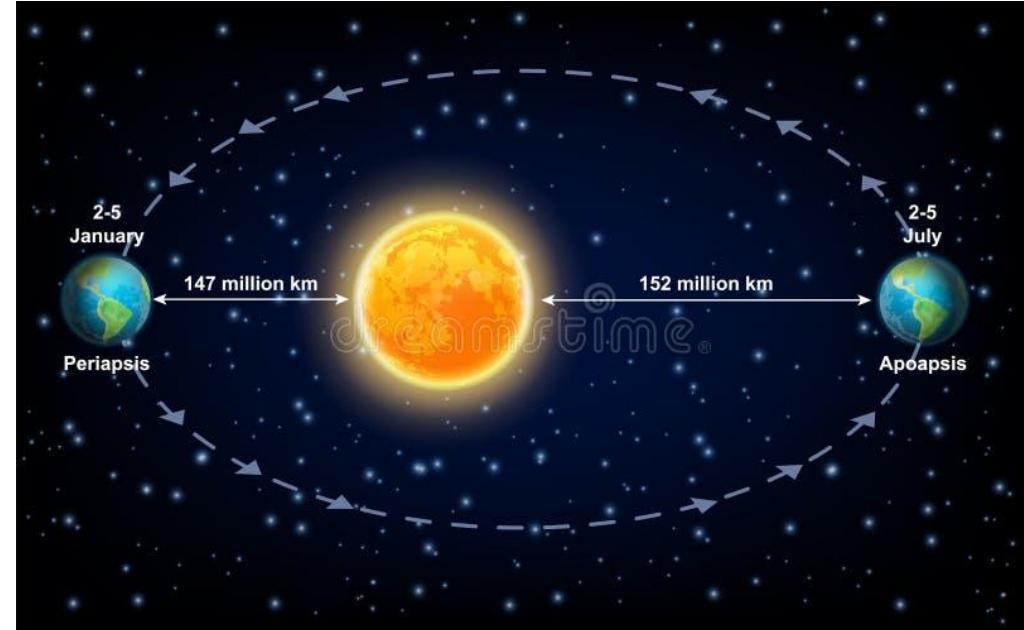
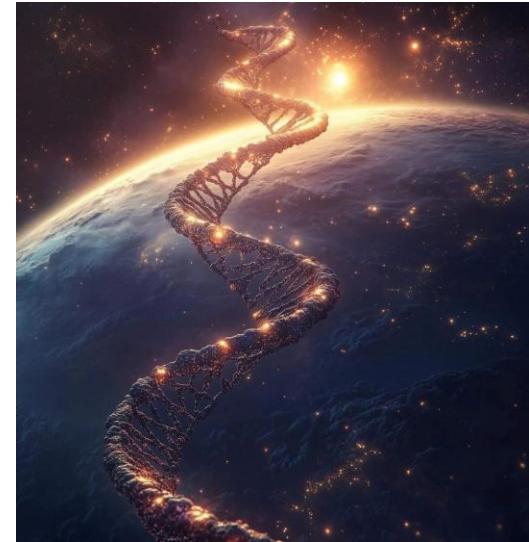
- Your DNA Could Go From Earth to the Sun 600 Times!
Your DNA is incredibly long - if stretched out, the DNA in your body could reach from the Earth to the Sun and back over 600 times.
- Each human cell contains approximately 6 feet of DNA, which is compacted into a structure called chromatin to fit inside the nucleus.
- If all the DNA in your body was uncoiled, it would stretch 67 billion miles long - equivalent to about 150,000 round trips to the Moon.

<https://www.ancestry.com/c/dna-learning-hub/dna-facts#:~:text=1.,stretch%20over%2067%20billion%20miles.>

A human cell contains 46 chromosomes:

The human body contains an estimated 37.2 trillion cells, with the number varying by sex and age:

- **Males:** Approximately 36 trillion cells
- **Females:** Approximately 28 trillion cells
- **Children:** Approximately 17 trillion cells
- **Newborns:** Approximately 26 billion cells



... starts at the field



Sampling



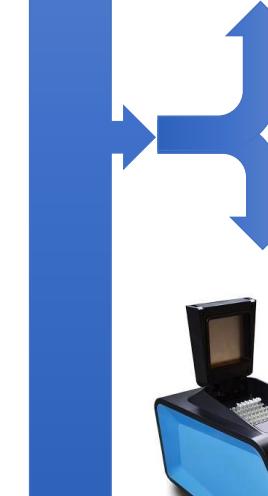
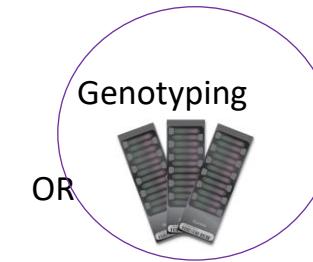
Extracting the DNA



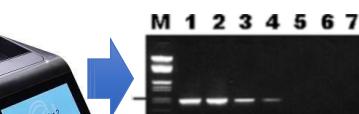
Quality check



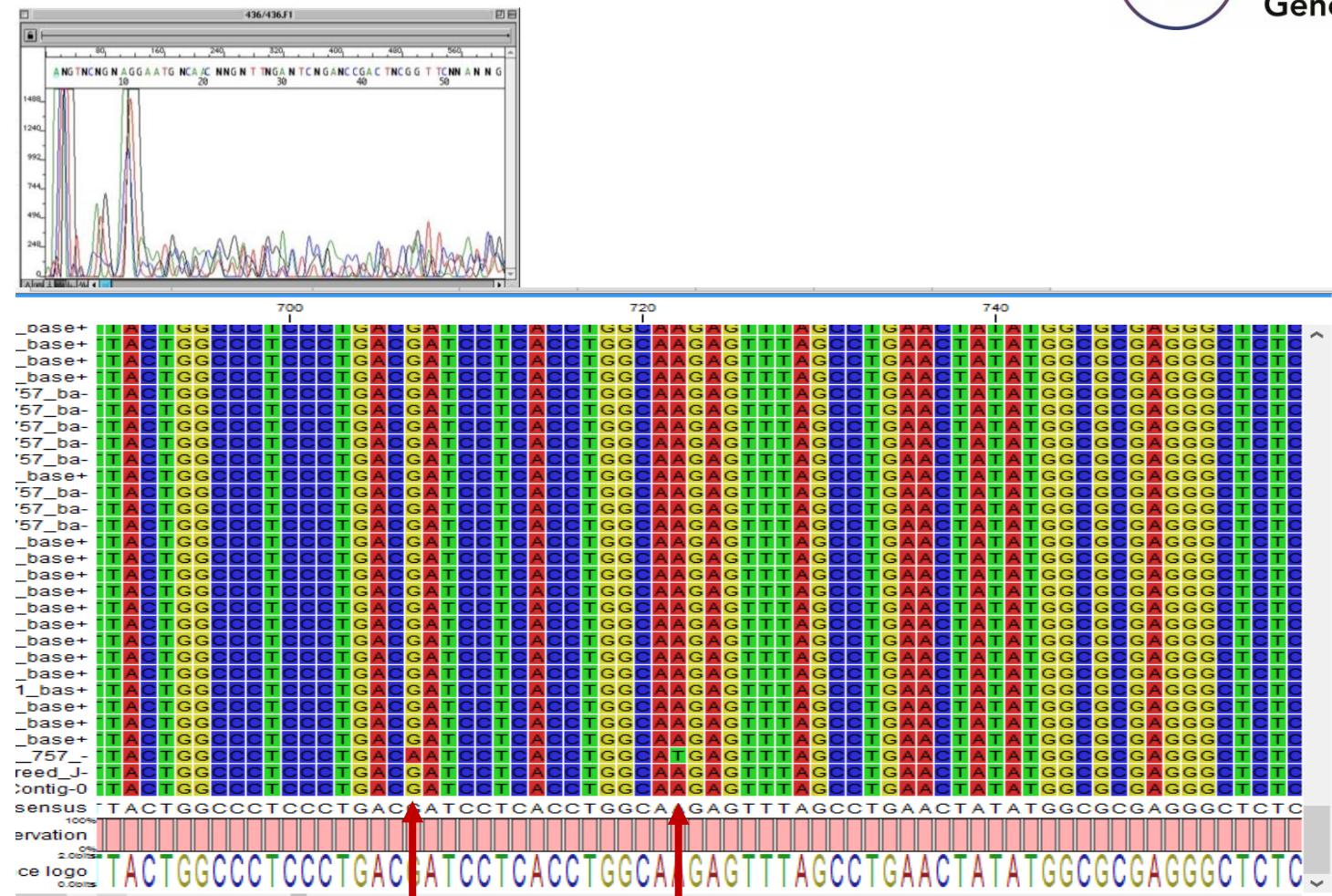
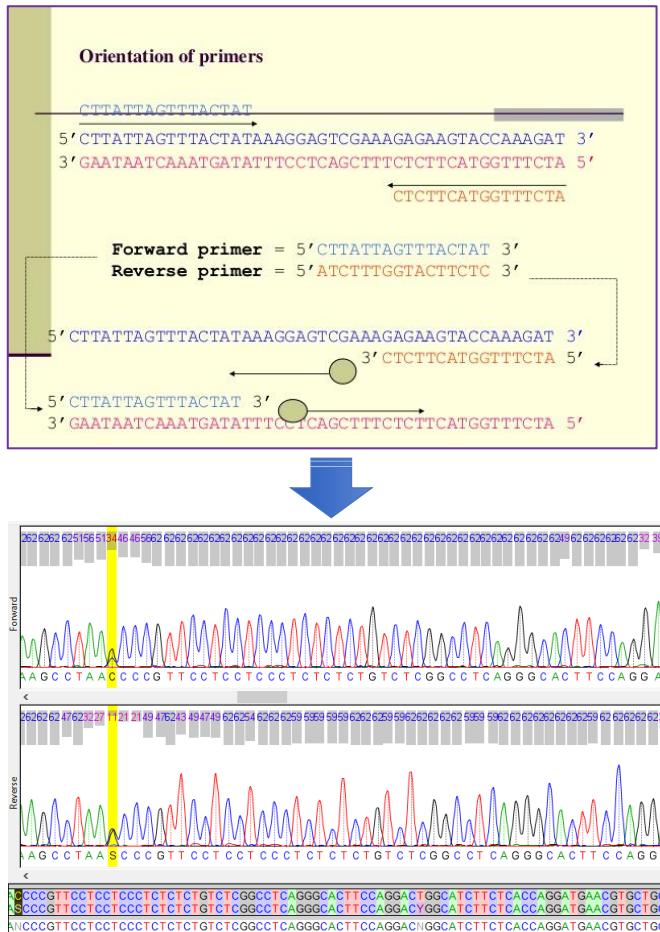
WGS



PCR



Sanger sequences



2 variants

- A chromatogram (sometimes also called electropherogram) is the visual representation of a DNA sample produced by a sequencing machine

WGS/NGS data

fastq format

```
@A00291:9:H5N3MDMX:1:1101:1181:1094 1:N:0:ATGCCTAA
GNTGGCTTGGGGTTTGGAATCGTGATACCAGAGGATGCCTACGAAAGAGTTAAATAC
+
F#FFFF:FFFF:::FFF:::FFFFFF:FFFFFF:FFFFFF:FFFFFF:FFFFFF:FFFFFF
```

Each read is represented by four lines. These lines are:

1. A sequence identifier with information about the sequencing run and the cluster.
2. The sequence (the base calls; A, C, T, G and N).
3. A separator, which is simply a plus (+) sign.
4. The base call quality scores.

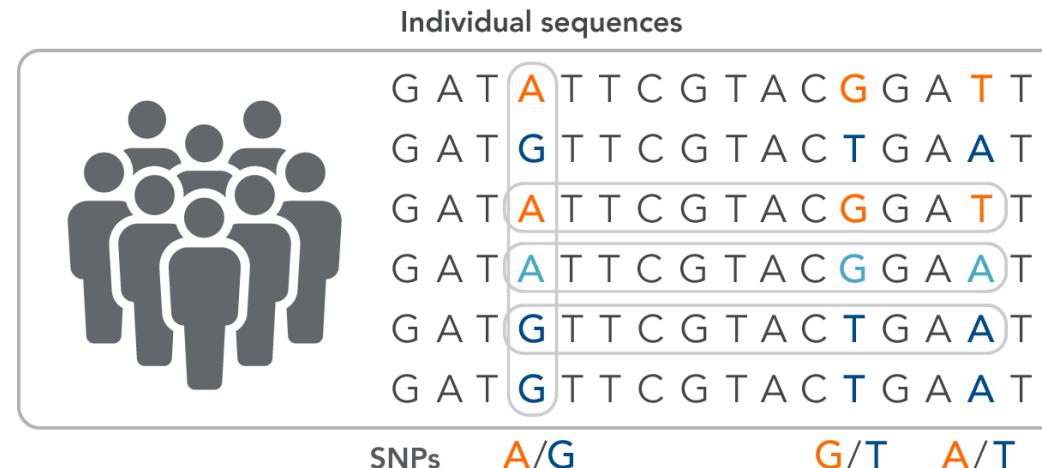
https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211



Genotype data

Single-nucleotide polymorphisms (SNP) genotypes

- A single nucleotide polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals.
- SNPs:
 - the most abundant genetic markers
 - have been widely used in **genomic** research
 - disease gene mapping
 - Medical and clinical diagnostics
 - forensic tests
 - genetic diversity
 - GWAS
 - Prediction



<https://eu.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know>



Are all mutations SNPs?

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century^{1,2} sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the first, second, third and fourth centuries of the history of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which genes are turned on and off. The fourth phase is the analysis of the mechanisms that regulate genes and with them the analysis of the mechanisms that regulate cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless race to decipher first genes and then entire genomes, spawning a new field of genomics. The fruits of this work already include the 3 sequences of over viruses and viroids, 205 naturally or artificially sequenced species, 31 bacteria, seven archaea, two animals and one plant.

Here we report the results of a collaboration involving 20 countries, the United States, the United Kingdom, Japan, Germany and China to produce a draft sequence of the genome. The draft sequence is now available online³ and includes a map covering more than 99% of the euchromatic part of the genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months since the start of the project. The sequence is continually updated daily throughout the project.⁴ The task ahead is to finish the sequence by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the bringing the vast majority of the sequence to this standard is straightforward and should proceed rapidly.

The analysis of the genome is of interest in its own right. It is the largest genome to be extensively sequenced, being 25 times as large as any previously sequenced genome, eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, the genome is now complete.

Much work remains to be done to produce a complete genome sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change

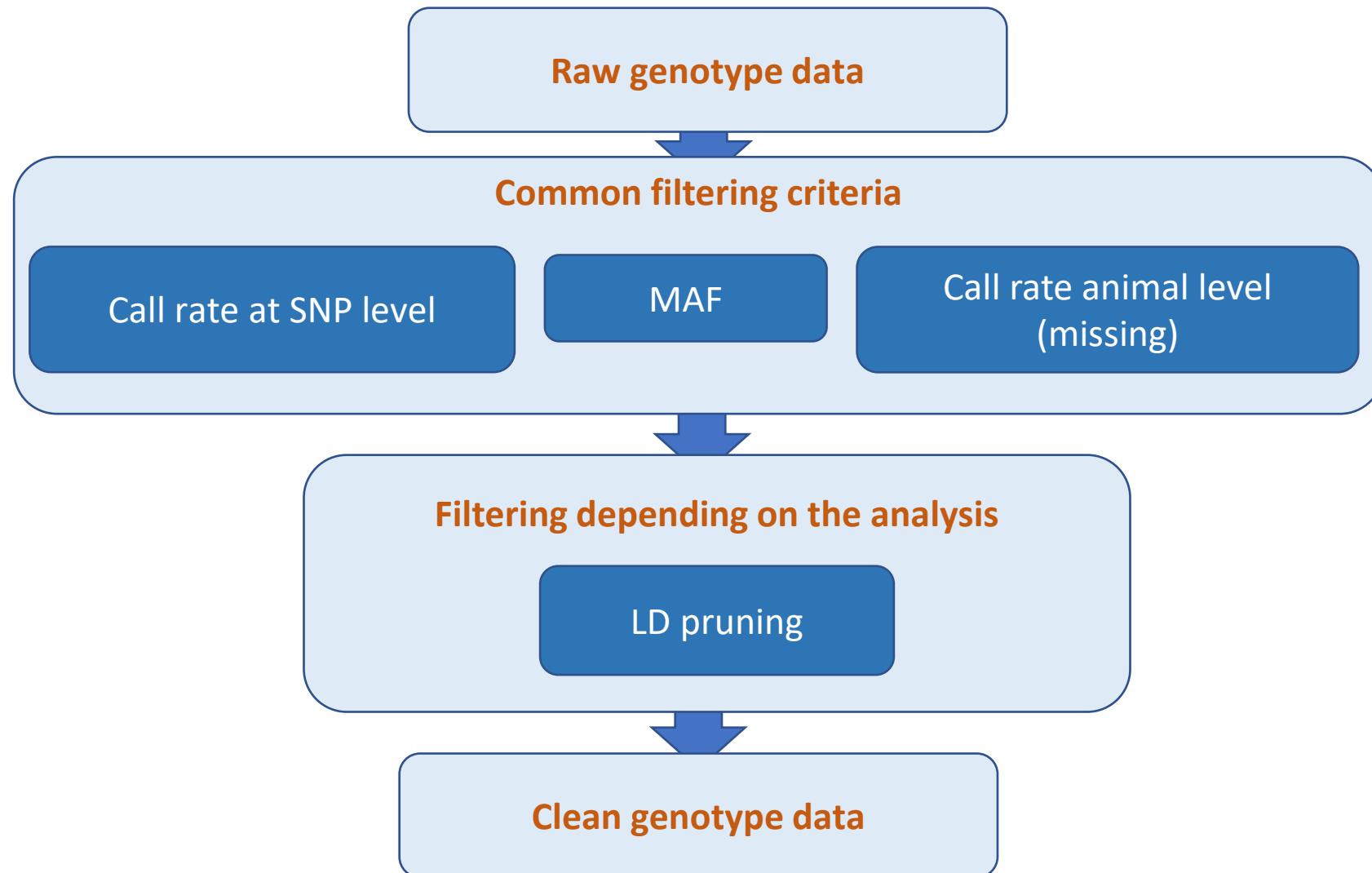
nature
the human genome

<http://neuroendoimmune.files.wordpress.com/2014/03/snp.png>

- To say a SNP, the available polymorphism needs to be observed at appreciable frequency (traditionally, **at least 1%**) in the human population for instance.



Quality control parameters



1. Call rate at SNP level: the call rate for a given SNP is **the proportion of individuals** in the study for which the corresponding SNP information is not missing.

- E.g. using a call rate of 95%, meaning we retain SNPs for which there is less than 5% missing data.

2. Sample-level filtering (call rate):

- Individuals who have missing genotype data across more than a pre-defined percentage of the typed SNPs need to be excluded.
- This proportion of missingness across SNPs is referred to as the **sample call rate**, and we apply a threshold of 95%.
- That is, individuals who are missing genotype data for more than 5% of the typed SNPs are removed.

	SNP	ANIMAL
025	1101011101S	111011111001000122110512051221102251111020122010210002211210200122010
036	21101110102201212222012101222010120222111112021222111112102020101101020111111200112110:	
050	12101002111202111200021212221000211221221221100000202200021102212221212202001112020:	
054	120001200220121211100121002222110211221102011212221200220021212121112021120220022100:	
066	20000202202102122112002200122221110122020110202222020220001222121011201021022010011010:	
097	10110212022012112211100210011111001022112120221111102022100120122012111021021012000:	
101	121100212022001122110001111220100101120112121211121201221002102002021211222022010022110:	
151	11100102022122021020110122020121221111221211112200221102011211110202200022102011220022102:	
172	21101202021111210121102110222010102112120221111102022100210220121110210211122022112011010:	
224	22000111022101221010102110252020111212022212221101021011102220050210121022010022125:	
277	210102200121221211212021012222002012210211020112102121002211010202111202101201010:	
314	122001112012220210201002121001120202012002012100202121211102022010101100:	
419	2211122101211202222102210211020102112121112200000011122002211122020222112120012121110:	
439	20020210012212121010102211010112202020221100101112100112010220122202102101011020:	
456	120001020221112200101021002211000202211212222001010221110210210112122111110211010:	
501	1110000212112120121212100222110102222110220011202110020211010202210002102022100021020:	
571	11000012022002212120220012100200111221101102222120022002120010102121202101010:	
579	112100210210010101110202202212000221112020222211022210120212112221111102101010:	
581	21100202100122120210010200201125121502252222502210112011210512022211211012110:	
657	11001112020111211101020012210001122121202112112002200222002212211212001112011:	
660	21000212022112022112101022211010122212221212111202012210122011211112111202200012101:	
730	2100020220020222200120022200122022220021102252200122001202111151001012022001012025:	
732	2121021215100220120001210112120121511021512252111502200111021110502022112201102010:	
764	1111021200122122110200012202012252221150220152222115022011020212005020202202211112110:	
780	121101021122220210101022002212012011212210121111011122102020200101122121200202102:	
800	22100012022122221020221101012112022102222200221002211121021202011022010111010:	
816	110001220220121220110022011121100011021121212002011222002221110211112102201020:	
832	121010011120011211100211122011112122212102011102022100211221001212111210211110:	
900	2101001102201221212110211021210221212110111102210012021211102110211021100220:	
901	12100102022112121221001002120201111221112122001111102210022010220122121201000120:	

3. MAF (SNP-level filtering): A large degree of homogeneity at a given SNP across study participants generally results in inadequate power to infer a statistically significant relationship between the SNP and the trait under study.

- remove SNPs for which the MAF is less than 1%.
 - In some instances, particularly small sample settings, a cut off point 5% is applied.
 - MAF is the lowest of the two allele frequencies
 - $p = freq(A)$
 - $q = 1 - p = freq(B)$
 - $MAF = \min(p, q)$
 - A fixed marker ($p = 0$ or $p = 1$) gives no information
 - An almost-fixed marker ($p = 0.0001$ or $p = 0.9999$) gives almost no info

4. LD pruning:

- Linkage disequilibrium (LD) is **the nonrandom association of alleles of different loci (Slatkin, 2008)**.
- **LD pruning** is the **removing loci based on high levels of pairwise LD**.
- Better results of **population structure** and **Principal Component Analysis (PCA)** are assumed to be obtained if the markers used are not in linkage disequilibrium with each other.
- If any pair of markers within the window are in LD greater than the specified threshold, the first marker in the pair will be pruned.



Centre for
Tropical Livestock
Genetics and Health

SNP genotype file format, and how can we manage the data?



Formats of the input file

0	snp6964-scaffold12561-13292	0	0
0	snp7949-scaffold12878-8765	0	0
0	snp7961-scaffold12886-28972	0	0
0	snp7962-scaffold1289-6273	0	0
0	snp8461-scaffold1305-10877	0	0
0	snp8463-scaffold1305-74425	0	0
0	snp8464-scaffold1305-146599	0	0
0	snp8466-scaffold1306-22563	0	0
0	snp8716-scaffold1310-23370	0	0
0	snp9138-scaffold1331-33874	0	0
0	snp9158-scaffold1333-23917	0	0
0	snp9659-scaffold13458-6636	0	0
0	snp9687-scaffold1348-7500	0	0
1	snp19065-scaffold1917-222828	0	0
1	snp4303-scaffold1134-142270	0	0
1	snp14078-scaffold1560-21647	0	21647
	Map file		
	snp14079-scaffold1560-51801	0	51801
1	snp14080-scaffold1560-100946	0	100946
1	snp14082-scaffold1560-185575	0	185575
1	snp14083-scaffold1560-228319	0	228319
1	snp14084-scaffold1560-297224	0	297224
1	snp14085-scaffold1560-330454	0	330454
1	snp14086-scaffold1560-378997	0	378997
1	snp14087-scaffold1560-412202	0	412202
1	snp14088-scaffold1560-452762	0	452762
1	snp14089-scaffold1560-488156	0	488156
1	snp14090-scaffold1560-516844	0	516844
1	snp14091-scaffold1560-581464	0	581464
1	snp14092-scaffold1560-614162	0	614162
1	snp14093-scaffold1560-666974	0	666974
1	snp14094-scaffold1560-721048	0	721048
1	snp14095-scaffold1560-755958	0	755958
1	snp14096-scaffold1560-786871	0	786871
1	snp14097-scaffold1560-830197	0	830197
1	snp14098-scaffold1560-872326	0	872326
1	snp14099-scaffold1560-920888	0	920888
1	snp14100-scaffold1560-986550	0	986550
1	snp14101-scaffold1560-1032913	0	1032913
1	snp2819-scaffold1082-727669	0	1087568
1	snp2817-scaffold1082-658683	0	1156554

AA	1011001	0	0
AB	1011002	0	0
AC	1011003	0	0
AD	1011004	0	0
AA	1011005	0	0
AB	1011006	0	0
AC	1011007	0	0
AD	1011008	0	0
AA	1011009	0	0
AB	1011011	0	0
AC	1011015	0	0
AD	1011016	0	0
AA	1011017	0	0
AB	1011018	0	0
AC	1011021	0	0
AD	1011022	0	0
AA	1011023	0	0
AB	1011026	0	0
AC	1011028	0	0
AD	1011029	0	0
AA	1011030	0	0
AB	1011032	0	0
AC	1011033	0	0
AD	1011035	0	0
AA	1011036	0	0
AB	1011037	0	0
AC	1011038	0	0
AD	1011039	0	0

Formats of the input file ... Cont'd

.bim, *.fam* and *.bed* file formats

Chr #	SNP Identifier	Genetic distance	Physical position SNP	Minor allele	Major allele	ABR	ET_ABR001	0	0	0	-9
0	snp6392-scaffold1227-41043	0	0	A	G	ABR	ET_ABR002	0	0	0	-9
0	snp6393-scaffold1227-75979	0	0	A	G	ABR	ET_ABR003	0	0	0	-9
0	snp6394-scaffold1227-146859	0	0	A	G	ABR	ET_ABR004	0	0	0	-9
0	snp6573-scaffold1233-677	0	0	A	G	ABR	ET_ABR005	0	0	0	-9
0	snp6621-scaffold1237-114663	0	0	G	A	ABR	ET_ABR006	0	0	0	-9
0	snp6643-scaffold1239-185	0	0	G	A	ABR	ET_ABR007	0	0	0	-9
0	snp6644-scaffold1239-42796	0	0	C	A	ABR	ET_ABR008	0	0	0	-9
0	snp6783-scaffold1249-6400	0	0	A	C	ABR	ET_ABR009	0	0	0	-9
0	snp6784-scaffold1249-39268	0	0	G	A	ABR	ET_ABR010	0	0	0	-9
0	snp6785-scaffold1249-70584	0	0	A	C	ABR	ET_ABR011	0	0	0	-9
0	snp6788-scaffold1249-185986	0	0	A	G	ABR	ET_ABR012	0	0	0	-9
0	snp6789-scaffold1249-220574	0	0	G	A	ABR	ET_ABR013	0	0	0	-9
0	snp6798-scaffold1249-261466	0	0	A	G	ABR	ET_ABR014	0	0	0	-9
0	snp6791-scaffold1249-290518	0	0	A	G	ABR	ET_ABR015	0	0	0	-9
0	snp6964-scaffold12561-13292	0	0	A	G	ABR	ET_ABR016	0	0	0	-9
0	snp7949-scaffold12878-8765	0	0	A	G	ABR	ET_ABR017	0	0	0	-9
0	snp7961-scaffold12886-28972	0	0	A	G	ABR	ET_ABR018	0	0	0	-9
0	snp7962-scaffold1289-6273	0	0	A	G	ABR	ET_ABR019	0	0	0	-9
0	snp8461-scaffold1305-10877	0	0	A	G	ABR	ET_ABR020	0	0	0	-9
0	snp8463-scaffold1305-74425	0	0	A	G	ABR	ET_ABR021	0	0	0	-9
0	snp8464-scaffold1305-146599	0	0	A	G	ABR	ET_ABR022	0	0	0	-9
0	snp8466-scaffold1306-22563	0	0	A	G	ABR	ET_ABR023	0	0	0	-9
0	snp8716-scaffold1310-23370	0	0	A	G	ABR	ET_ABR024	0	0	0	-9
0	snp9138-scaffold1331-33874	0	0	A	C	ABR	ET_ABR025	0	0	0	-9
0	snp9158-scaffold1333-23917	0	0	A	G	ABR	ET_ABR026	0	0	0	-9
0	snp9659-scaffold13458-6636	0	0	A	C	ABR	ET_ABR027	0	0	0	-9
0	snp9687-scaffold1348-7500	0	0	A	C	ABR	ET_ABR028	0	0	0	-9
1	snp19065-scaffold1917-222828	0	0	A	G	ABR	ET_ABR029	0	0	0	-9
1	snp4303-scaffold1134-142270	0	0	G	A	ABR	ET_ABR030	0	0	0	-9
1	snp14078-scaffold1568-21647	0	21647	A	C	ABR	ET_ABR031	0	0	0	-9
1	snp14079-scaffold1568-51801	0	51801	A	G	ABR	ET_ABR032	0	0	0	-9
1	snp14080-scaffold1568-100946	0	100946	A	G	ABR	ET_ABR033	0	0	0	-9
1	snp14082-scaffold1568-185575	0	185575	A	G	ABR	ET_ABR034	0	0	0	-9
1	snp14083-scaffold1568-228319	0	228319	C	A	ABR	ET_ABR035	0	0	0	-9
1	snp14084-scaffold1568-297224	0	297224	A	G	ABR	ET_ABR036	0	0	0	-9
1	snp14085-scaffold1568-330454	0	330454	A	G	ABR	ET_ABR037	0	0	0	-9
1	snp14086-scaffold1568-378997	0	378997	G	A	ABR	ET_ABR038	0	0	0	-9
1	snp14087-scaffold1560-412202	0	412202	A	G	ABR	ET_ABR039	0	0	0	-9
			452762	A	G						+

- Major allele: the most common allele for a given SNP
 - Minor allele: the least common (or rarer) allele



File formats for GBLUP analysis

UGA42014 210021212111121110010011121102001210220012120101101211110211211210101101110102
UGA42019 20101120211211201011112012100011012111210021202021102121121112220210000110102
UGA42029 10111020112111010000112112110010111102200121111121001112120121211120110111021001
UGA42039 100020101221022010101112001101020111102111011202111112121121122101211110111121111
UGA42047 20002020222202200000000200220020022002220202220202220200000020002
UGA42051 200020202222022000000002002200200220022202022002220202220200000020002
UGA42052 200020202222022000000002002200200220022202022002220202220200000020002
UGA42056 1021011100212100210001111221200111011011001210211200011111020210121021111012110
UGA42057 200020202222022000000002002200200220022202022002220202220200000020002
UGA42061 100020101221022010101112001101020111102111011202111112121122101211110111121111
UGA42085 101011201121022101010002112201111121122101212011211122112022120210201001120012
UGA42088 00111010012011102010122211101101020020211110111210111011221021110111020222122110
UGA42094 01121110111110110000221222121010111212110210011210111002211011121020010211121101
UGA42095 2000202022220220000000020022002002200220022002200220022202022202220200000020002
UGA42098 10002010122102201010111200110102011110211101120211112121122101211110111121111
UGA42101 01210120101111012001021212102001121121110111100011220011210020120111011110
UGA42108 200020202222022000000002002200200220022202022202220200000020002
UGA42109 101110201121111010000112112100101111022001211111210011121211120110111021001
UGA42127 20002020222202200000000200220020022002200220022202022202220200000020002
UGA42136 100020101221022010101112001101020111102111011202111112121122101211110111121111
UGA42137 10210111002121002100011112212001110110110012102112000111110202101210211110121111
UGA42138 01210120101111012001021212102001121121110111100011220011210020120111011110
UGA42139 1000201012210220101011120011010201111021101202111112121121122101211110111121111

SNP data file

Index	Name	Chromosome	Position	GenTrain	Score	SNP	ILMN	Strand	Customer	Strand	NormID
1	ARS-BFGL-BAC-10172	14	6371334	0.9176	[A/G]	TOP	TOP	2			
2	ARS-BFGL-BAC-1020	14	7928189	0.9413	[T/C]	BOT	TOP	2			
3	ARS-BFGL-BAC-10245	14	31819743	0.7646	[T/C]	BOT	BOT	2			
4	ARS-BFGL-BAC-10345	14	6133529	0.8906	[A/C]	TOP	TOP	2			
5	ARS-BFGL-BAC-10365	14	27005721	0.9206	[A/C]	TOP	TOP	BOT			1
6	ARS-BFGL-BAC-10375	14	6616434	0.9258	[A/G]	TOP	TOP	2			
7	ARS-BFGL-BAC-10591	14	17544926	0.7439	[A/G]	TOP	TOP	1			
8	ARS-BFGL-BAC-10867	14	34639444	0.9085	[G/C]	BOT	BOT	101			
9	ARS-BFGL-BAC-10919	14	31267746	0.8255	[A/G]	TOP	TOP	2			
10	ARS-BFGL-BAC-10951	10	17911906	0.9056	[T/G]	BOT	BOT	2			
11	ARS-BFGL-BAC-10952	10	18882288	0.9184	[A/G]	TOP	TOP	2			
12	ARS-BFGL-BAC-10960	10	20609250	0.5678	[A/G]	TOP	TOP	2			
13	ARS-BFGL-BAC-10972	10	20792754	0.8432	[G/C]	BOT	BOT	102			
14	ARS-BFGL-BAC-10975	10	21225382	0.7991	[A/G]	TOP	TOP	2			
15	ARS-BFGL-BAC-10986	10	26527257	0.8941	[A/C]	TOP	BOT	2			
16	ARS-BFGL-BAC-10993	10	78512500	0.8649	[A/G]	TOP	BOT	2			
17	ARS-BFGL-BAC-11000	10	79252023	0.9433	[T/G]	BOT	BOT	2			
18	ARS-BFGL-BAC-11003	10	80410977	0.8842	[T/C]	BOT	BOT	2			
19	ARS-BFGL-BAC-11007	10	80783719	0.9110	[T/C]	BOT	BOT	2			
20	ARS-BFGL-BAC-11025	10	84516867	0.8711	[T/G]	BOT	BOT	2			

Map file

Quality control QC using plink in R

The following script helps to evaluate the QC in plink.

- Download plink from: <https://www.cog-genomics.org/plink/>
- put plink (only the executable file) at your working directory

Quality Control exercise

```
####convert the data from .ped to .bed format
system("./plink --file Sheep05_724 --make-bed --chr-set 26 --out Sheep05_724")
#missing at SNP level
system("./plink --bfile Sheep05_724 --chr-set 26 --make-bed --geno 0.01 --out Sheep05_afterQC")
#minor allele frequency
system("./plink --bfile Sheep05_724 --chr-set 26 --make-bed --geno 0.10 --maf 0.10 --out Sheep05_afterQC")
#missing at animal level
system("./plink --bfile Sheep05_724 --chr-set 26 --make-bed --geno 0.10 --maf 0.10 --mind 0.05 --out
Sheep05_afterQC")
```



Id-pruning

LD-pruning:

```
system("./plink --bfile Sheep05_afterQC --chr-set 26 --indep-pairwise 50 10 0.5 ")
system("./plink --bfile Sheep05_afterQC --extract plink.prune.in --make-bed --chr-set 26 --out
Sheep05_afterQC_pruned")
```

#To generate the PCA:

```
system("./plink --bfile Sheep05_afterQC_pruned --chr-set 26 --pca --outSheep05_afterQC_pruned_pca ")
```

#to generate structure format

```
system("./plink --bfile Sheep05_afterQC_pruned --recode-structure --chr-set 26 --out
Sheep05_afterQC_pruned_structure")
```

Note: For the admixture, you can still use the .bed format

Tips

#To generate other population genetics parameters:

#To generate the MAF from bed file:

```
system("./plink --bfile Sheep05_afterQC --chr-set 26 --freq --out Sheep05_afterQC_maf")
```

To generated expected and observed heterozygosity, and hwe:

```
system("./plink --bfile Sheep05_afterQC --hardy --chr-set 26 --out Sheep05_afterQC_het")
```

#To generate coefficient of inbreeding:

```
system("./plink --bfile Sheep05_afterQC --chr-set 26 --het --out Sheep05_afterQC_Fx")
```

Why checking the PCA is important before proceeding to genomic evaluation?

- It helps to know the structure of the target population so that it gives idea whether we need to implement stratification in the model we will be fitting
- It also helps to provide feedback to the farmers about the structure of the animals they own

PCA

```

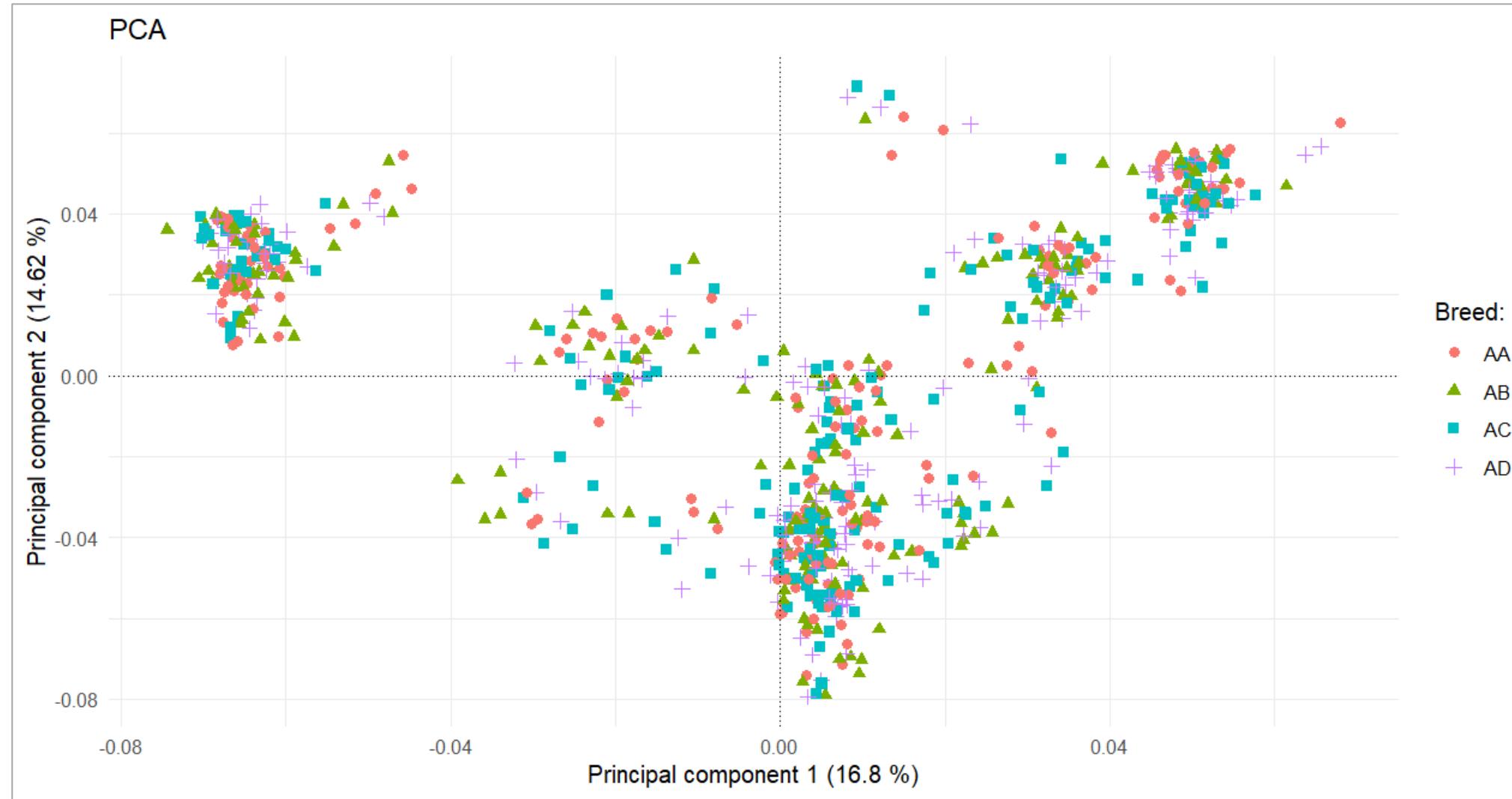
install.packages("tidyverse")
library(tidyverse)
# read in result files
eigenValues <- read_delim("PATH/Sheep05_afterQC_pruned_pca.eigenval", delim = " ", col_names = F)
eigenVectors <- read_delim("PATH/Sheep05_afterQC_pruned_pca.eigenvec", delim = " ", col_names = F)
eigenVectors
eigenValues
## Proportion of variation captured by each vector
eigen_percent <- round((eigenValues / (sum(eigenValues)))*100), 2)
# PCA plot
ggplot(data = eigenVectors) +
  geom_point(mapping = aes(x = X3, y = X4, color = X1, shape = X1), size = 1, show.legend = F ) +
  geom_hline(yintercept = 0, linetype="dotted") +
  geom_vline(xintercept = 0, linetype="dotted") +
  labs(title = "PCA",
       x = paste0("Principal component 1 (",eigen_percent[1,1]," %)"),
       y = paste0("Principal component 2 (",eigen_percent[2,1]," %)"),
       colour = "Breed:", shape = "Breed:") +
  theme_minimal()

```

PCA plot generated from the example data (*blackface*)



Centre for
Tropical Livestock
Genetics and Health



Admixture package

Why admixture is preferred?

([Alexander and Lange., 2012. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. Read](#))

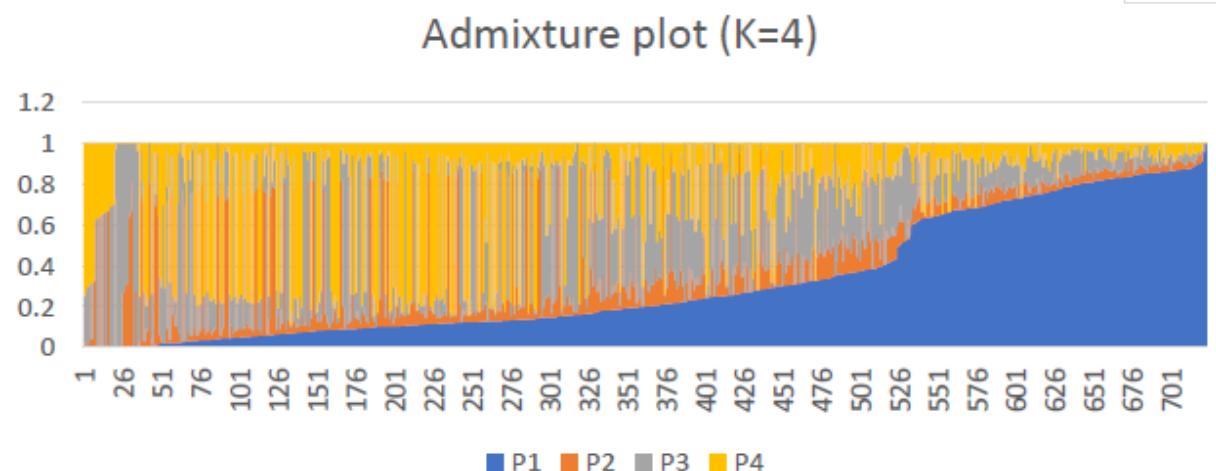
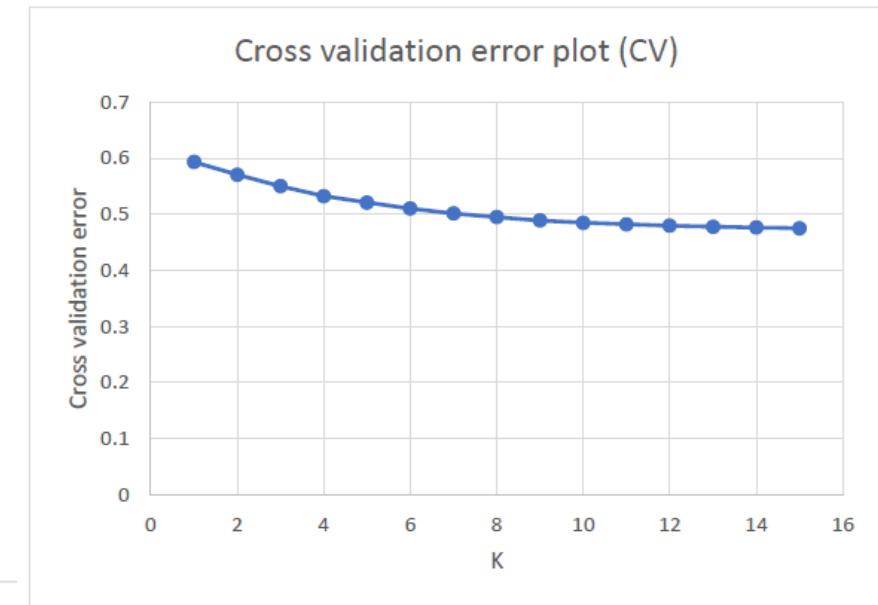
- *Helps to generate the breed proportion of large dataset*
- ADMIXTURE estimates individual ancestries by efficiently computing maximum likelihood estimates in a parametric model.
 1. ADMIXTURE can be used to estimate the number of underlying populations through **cross-validation**.
 2. Individuals of known ancestry can be exploited in supervised learning to yield more precise ancestry estimates.
 3. By penalizing small admixture coefficients for each individual, one can encourage model parsimony, often yielding more interpretable results for small data sets or data sets with large numbers of ancestral populations.
 4. By exploiting multiple processors, large data sets can be analyzed even more rapidly.

Running admixture

- **Generating the input file:**
- ADMIXTURE requires unlinked (i.e.LD-pruned) SNPs in plink format.
- It is very easy to generate the input file from a VCF containing such SNPs.
- Plink helps to generate the .bed file which can be read by ADMIXTURE
- The default cross-validation 5-fold CV (you can change as the K value you expect to be) and starts as K=2.
- **ADMIXTURE produced 2 files:** [Q](#) which contains cluster assignments for each individual and. [P](#) which contains for each SNP the population allele frequencies.
- The default cross-validation 5-fold CV (you can change as the K-value you expect to be) and starts as K = 2.
- Grep the results to access the CV error log Files as: [grep 'CV error' log_*](#)

- Running population admixture in using admixture package

```
bed=/PATH/shp724afterQCFinal.bed
mkdir -p / PATH /newFolder
out=/ PATH / newFolder
cd / PATH / newFolder
for k in $(echo {1..15..1});
do echo ${k};
./admixture --cv=10 ${bed} ${k} > ${out}/log_${k}.txt;
done
grep 'CV error' log_*
```



Formatting the genotype data for GWAS

- # pick chromosomes of interest for GWAS:
- `./plink --file Sheep05_724 --recode --chr 1-26 --out Sheep05_afterQC`
- #converting from Allele form to numeric for blupf90 package
- `./plink --bfile Sheep05_afterQC --chr-set 26 --recode A --out shp_blupf90`
- # removing first line or header
- `cat shp_blupf90.raw | sed 1d > snpblup.txt`
- #rmoving column
- `awk '{$1=$3=$4=$5=$6="";print $0}' snpblup.txt > snpblup1.txt`
- #removing space from column
- `awk '{s=$1;gsub($1 FS,x);$1=$1;print s FS $0}' OFS= snpblup1.txt > snpblup2.txt`
 - OR use this command: `sed -i "s/ //g" snpblup1.txt`
- #adding and removing space between column 1 and 2
- `awk '{print ""$1" "$2 }' snpblup2.txt >snpblupclean3.txt`
- #slecting map file
- `awk '{print ""$2" "$1" "$4 }' shp724afterQCFinal.map > gwasmapxxx.txt`



Centre for
Tropical Livestock
Genetics and Health

CTLGH Funders

BILL & MELINDA
GATES foundation



Biotechnology and
Biological Sciences
Research Council

