NGS data - Workflow, formats and programs

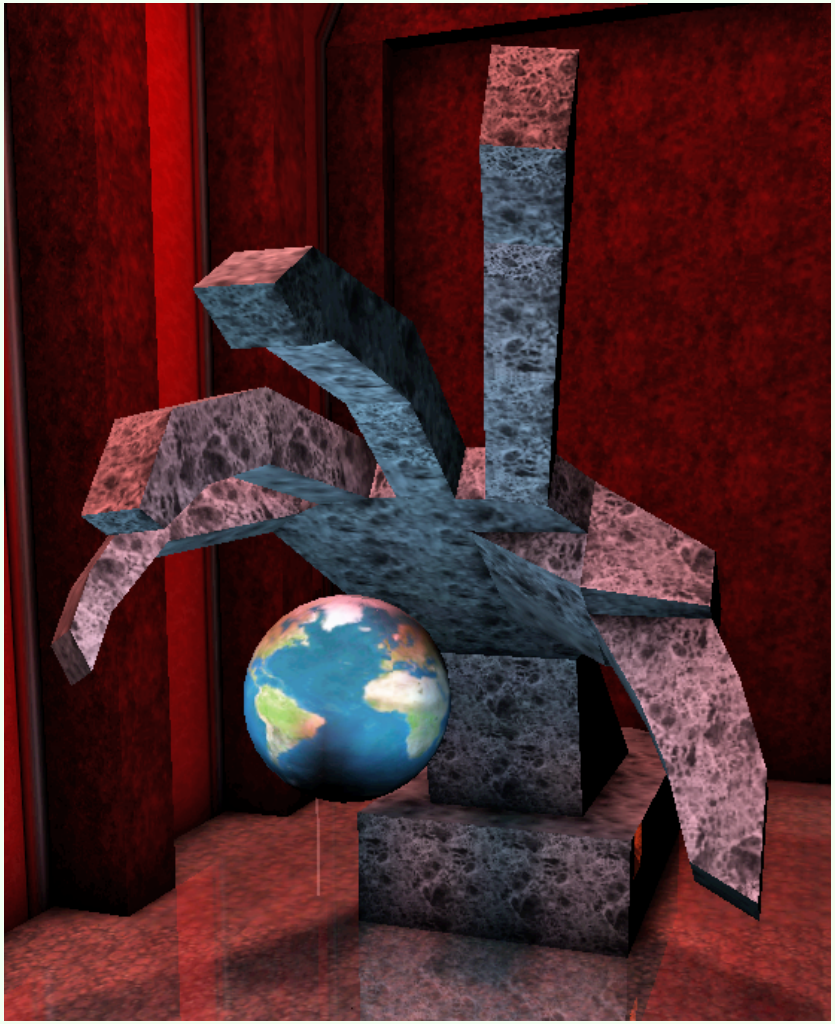
Thorfinn Sand Korneliussen
The GLOBE institute, University of Copenhagen

August 30, 2023

Outline

1. Background(GLOBE) and Learning objectives
2. Ancient DNA introduction
3. Low level formats
   - data types
   - Fasta and fastq files
   - Computer exercise
4. Alignment and variant files
   - SAMfiles
   - Concepts and definitions
   - VCF files
   - Computer exercises

GLOBE Institute



- New institute
- Five sections
- Lundbeck Center for geogenetics
- Eight groups
- Korneliussen Group:
  Rasmus Henriksen (tomorrow)
  Abigail Ramsoe (Friday)
  Isin Altinkya (computer exercises)

Learning objectives

Week1

Monday NGS intro

Tuesday Aligning data to a reference genome

Friday Modelling uncertainty of NGS data

After this week you will be able to navigate all commonly used files and software used in the context of NGS data. You will have a through understanding of how modern aligners work and will be able to apply this. NGS data is associated with high error rates and you will learn the theory of how this is dealt with.

Today

- Fileformats used in NGS data (fasta,fq,sam,vcf)
- Standard tools NGS projects(samtools,bcftools)
- Introduction to important concepts (depth,genotype, variable sites)

Standard workflow

- Lab stuff + sequencing
- Call bases and quality scores (FastQ)
- Alignment or *denovo* assembly. Alignment is topic for tomorrow

Genotypes

- Generate genotype likelihoods. Uncertainty is topic for friday.
- Estimate allele or haplotype frequencies
- Calculate genotype probabilities
- Call genotypes

Downstream analysis

- Variation detection
- Population genetic analysis

Ancient DNA introduction

After this sub-session you will:

- Be able to describe the characteristics of aDNA and the difficulties in studying aDNA.
- Understand important concepts (e.g. library, sequencing, complexity, clonality, DoC)
- Be able to participate in a meaningful discussion of aDNA projects and literature.

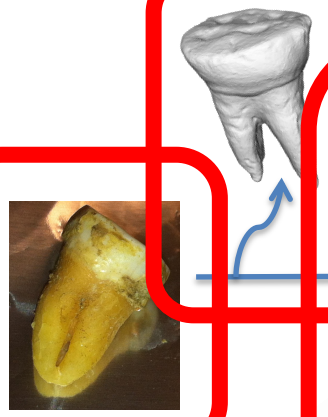
Sample types

- Bone
- Teeth, calculus(dental)
- Hair
- Fur, skin
- Shells
- Plants/seed/pollen
- Coprolites (Fossil feces)
- Sediments
- Ice cores
- Coral?

Sample collection

DNA wet lab work

Data analysis



Binding buffer

Digestion buffer

Elution

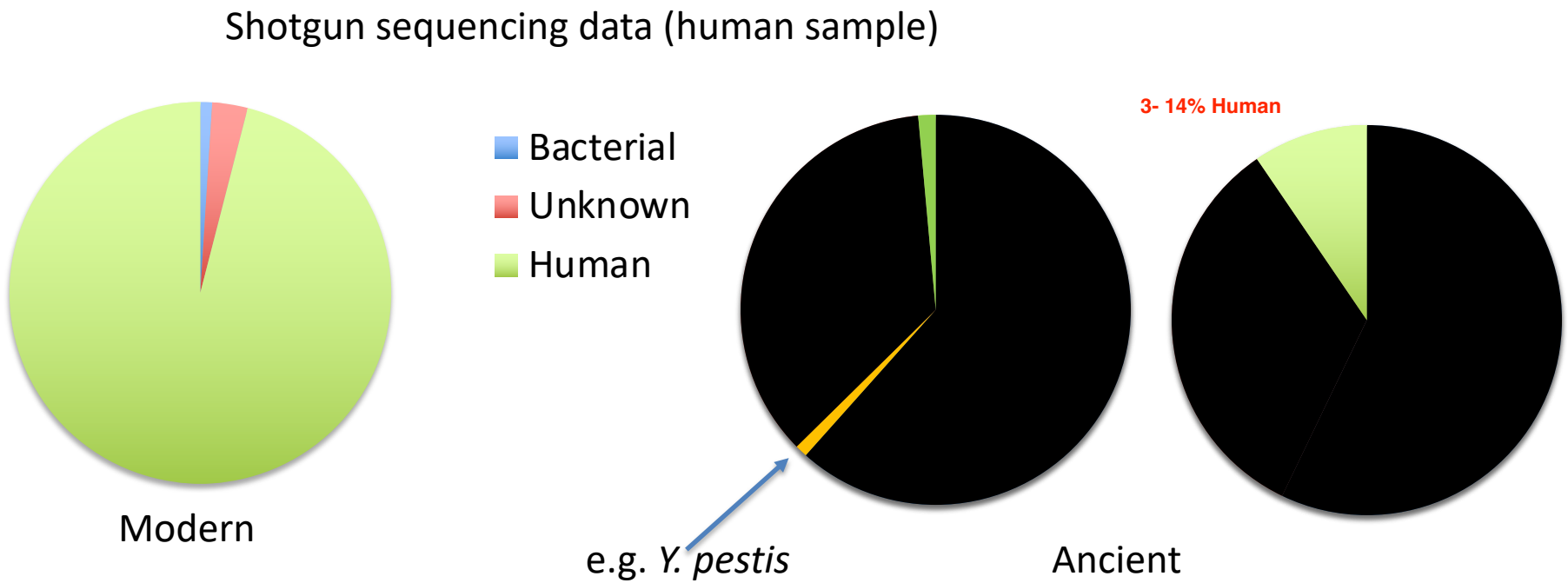
ancient DNA (aDNA)

Enzymes + Adapters

Buffers + dNTPs...

Sequencing library

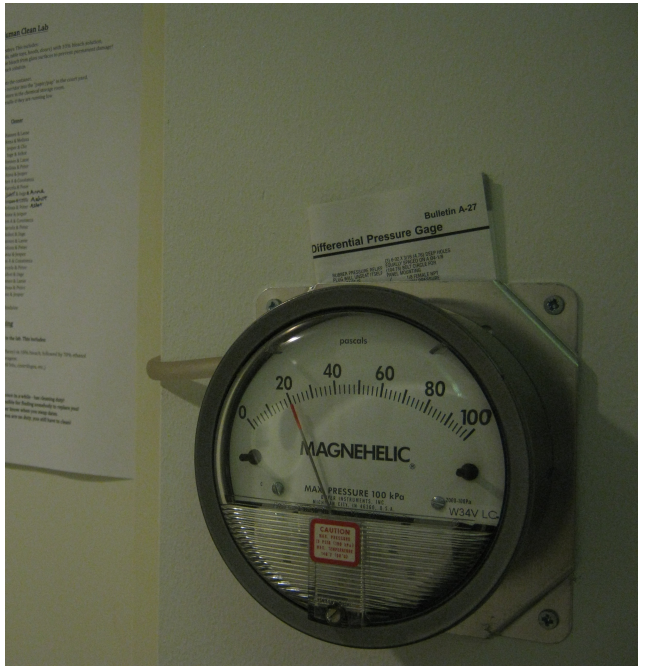
Low endogenous [DNA] in most ancient samples



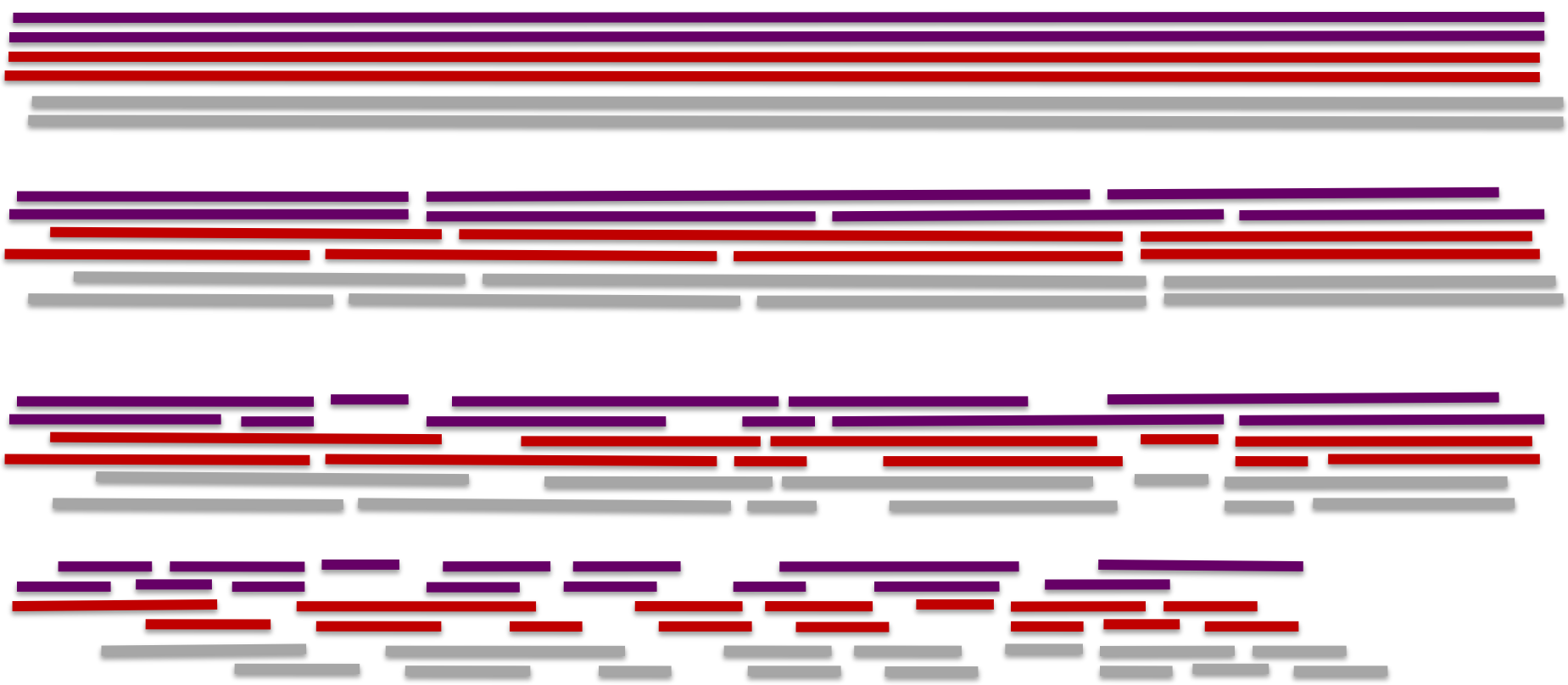
Avoid contamination with modern human DNA by...

- Minimize contact at sample collection
  - Gloves,
  - Zip-lock bags,
  - (Face mask)
  - Minimal cleaning of specimens,
  - Refrigerated storage (-20°C).
- Apply aDNA lab procedures
- Track contamination
  - mtDNA
  - Y-chromosome
  - X-chromosome

Ancient DNA lab at Centre for GeoGenetics



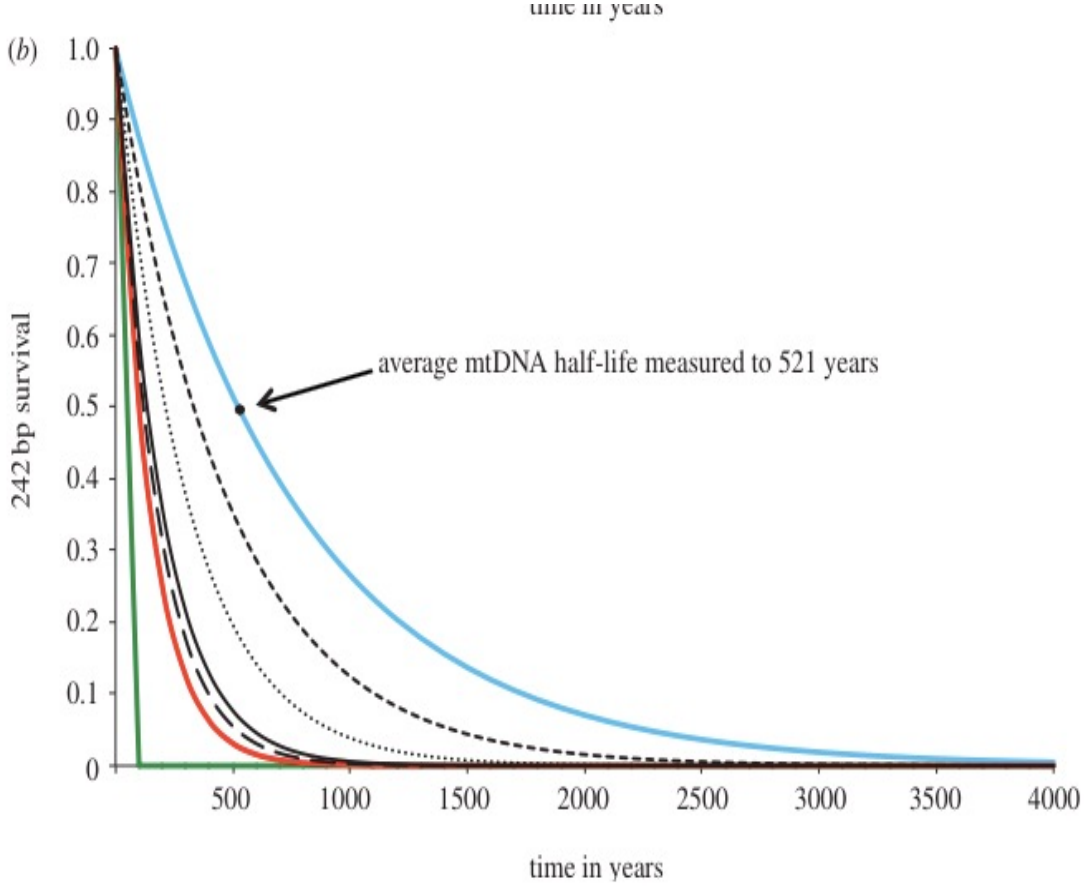
DNA breaks down post mortem

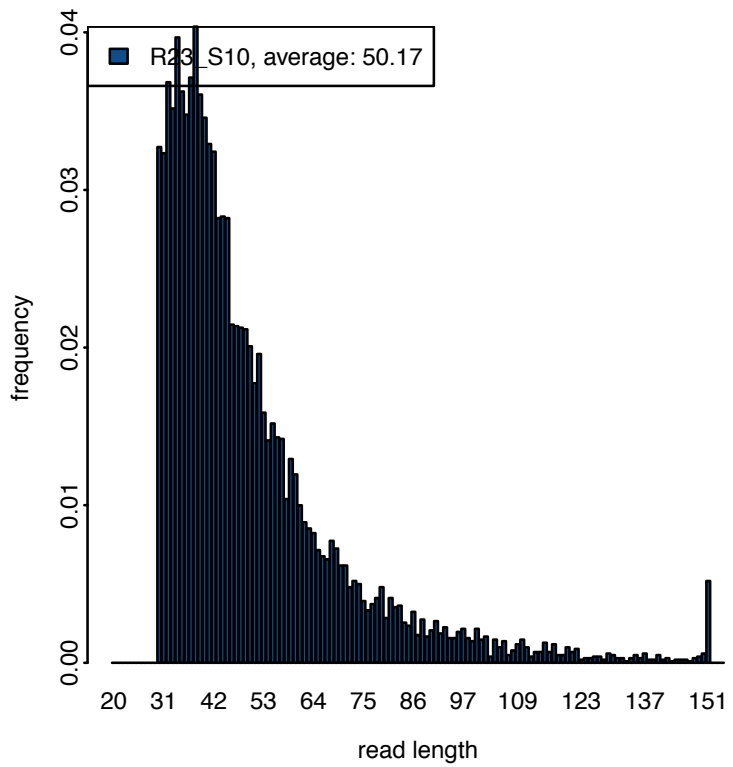


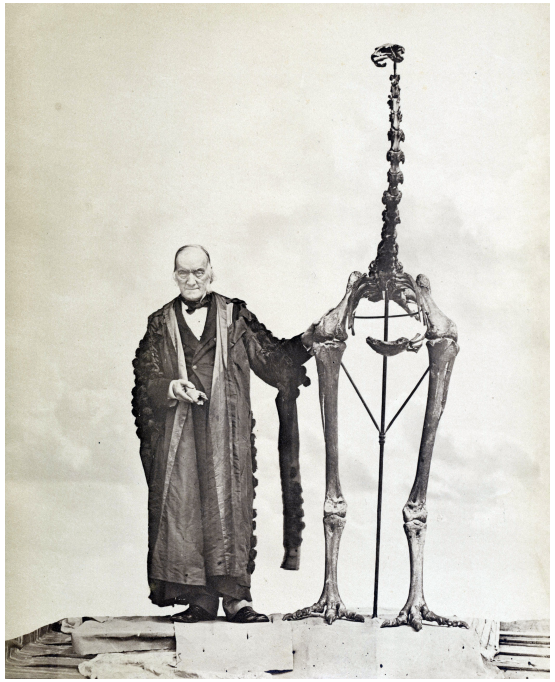
Time

at leas 20bp, otherwise the error just like a random mapping

DNA breaks down post mortem







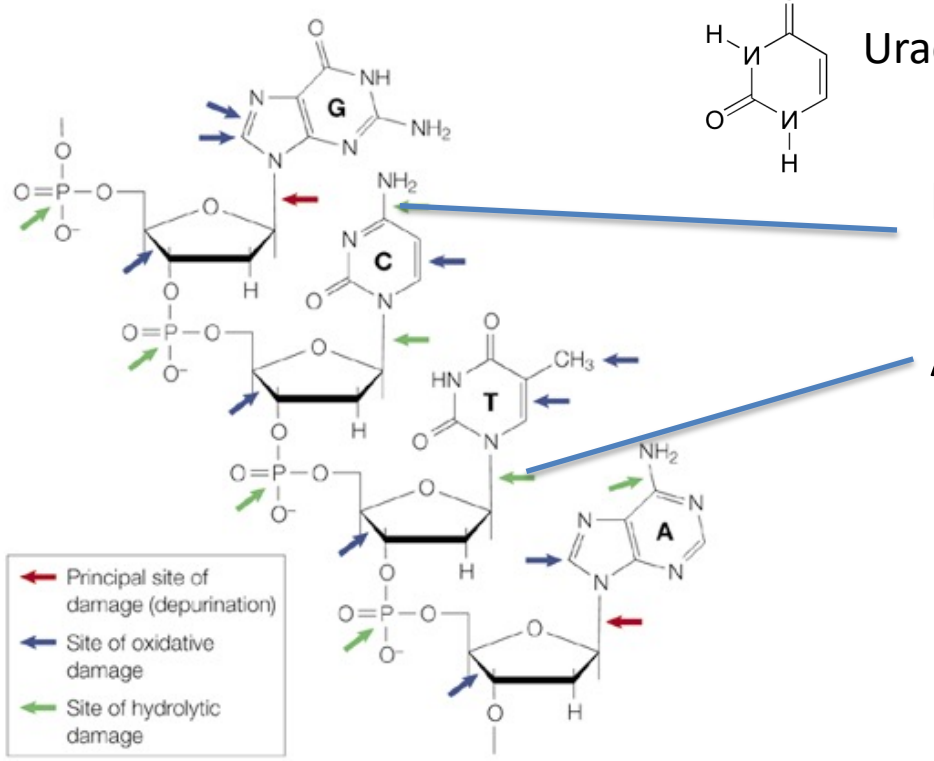
Allentoft 2012

Additional reading: Kistler NAR 2017

DNA **decay** and **damage**

DNA damage results in

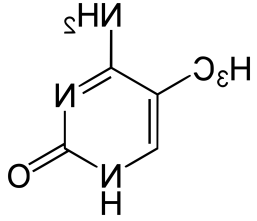
- Short DNA lengths (break of backbone)
- Post mortem change or loss of bases.



Uracil (U)

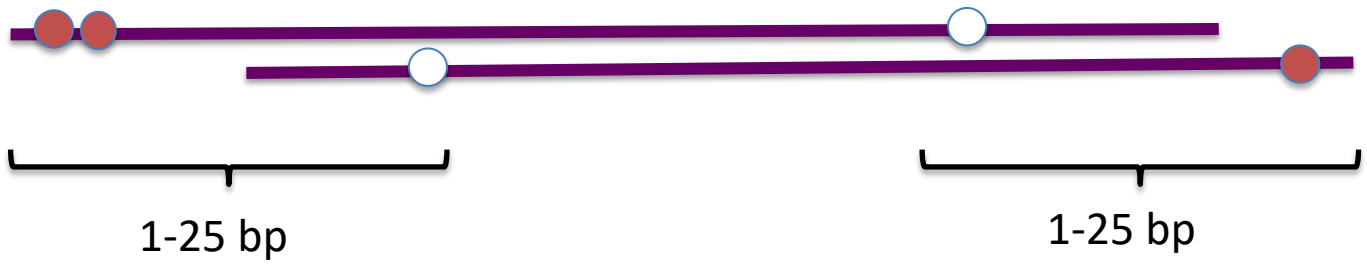
Deamination (C -> U)

Abasic sites

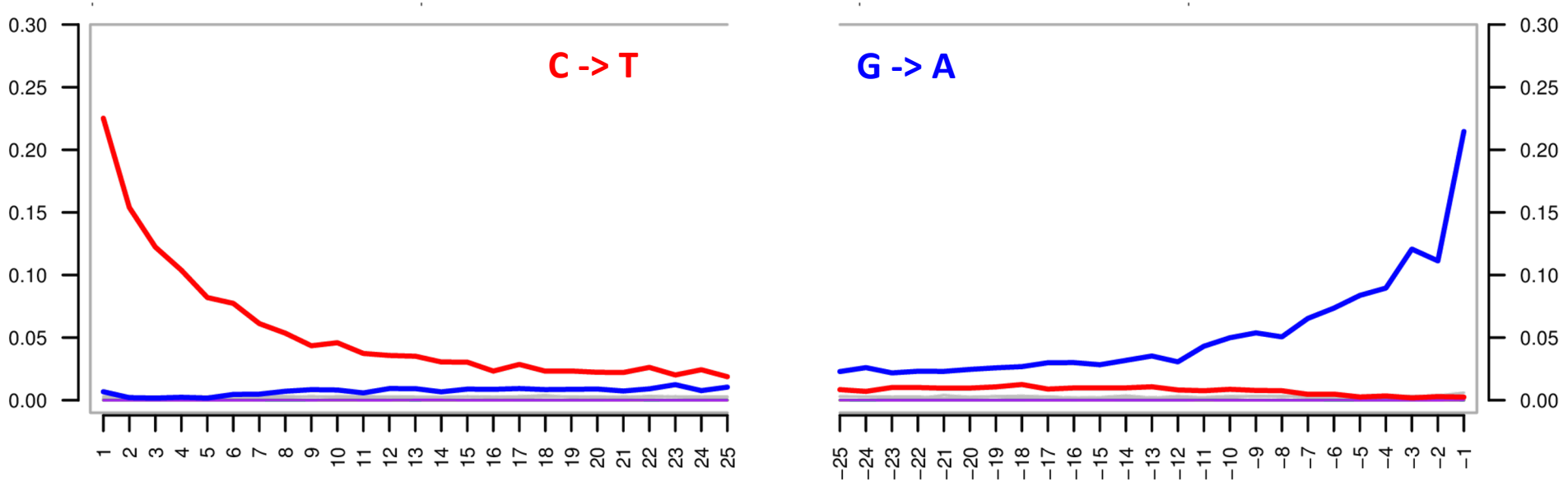


Nature Reviews | Genetics

aDNA is damaged

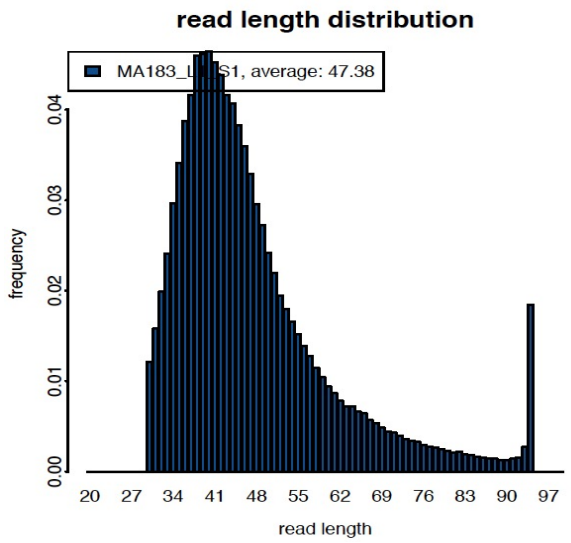


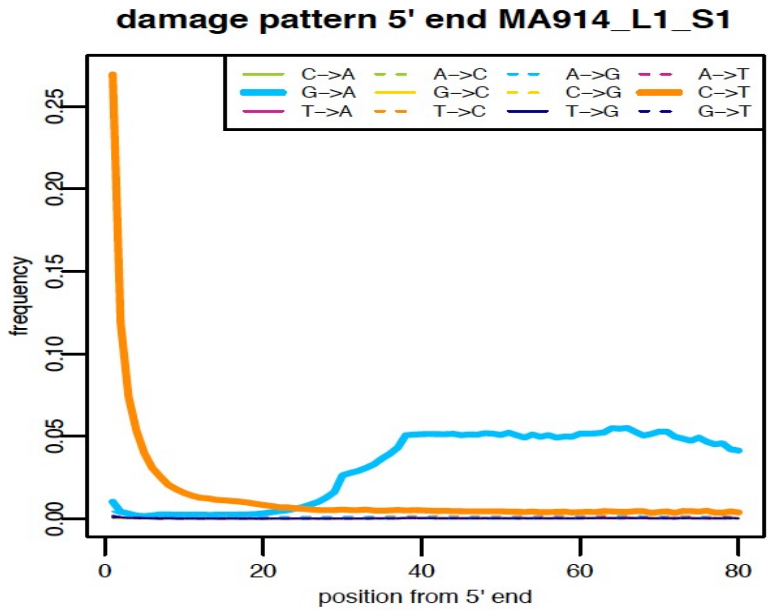
Ginolhac et al.



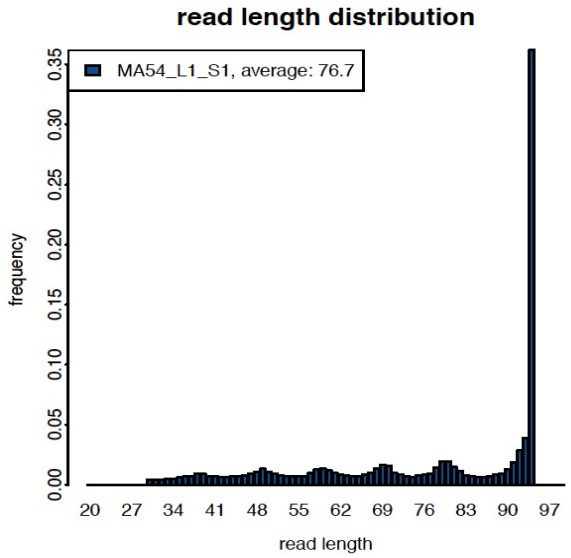
Which sample is ancient?

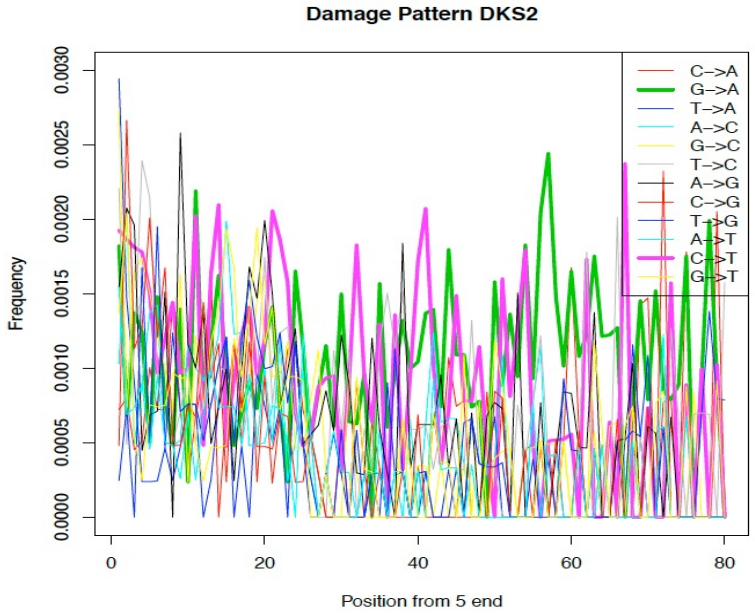
1





2

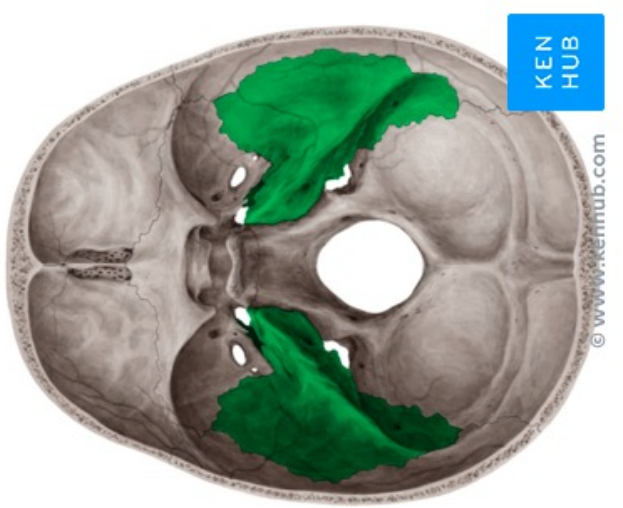




Degradation of DNA

1. What factors affects the rate of decay? Exposure to sunlight, temperature, microbes
2. How can the damage be useful? Validate whether it is ancient or not e.g. C->T damage, or G->A damage.
3. In what locations/regions/environments would you expect good samples? Dry and cold
4. Where would you predict problems?
5. How can we analyse the data in presence of damage? Text

Bone types matter



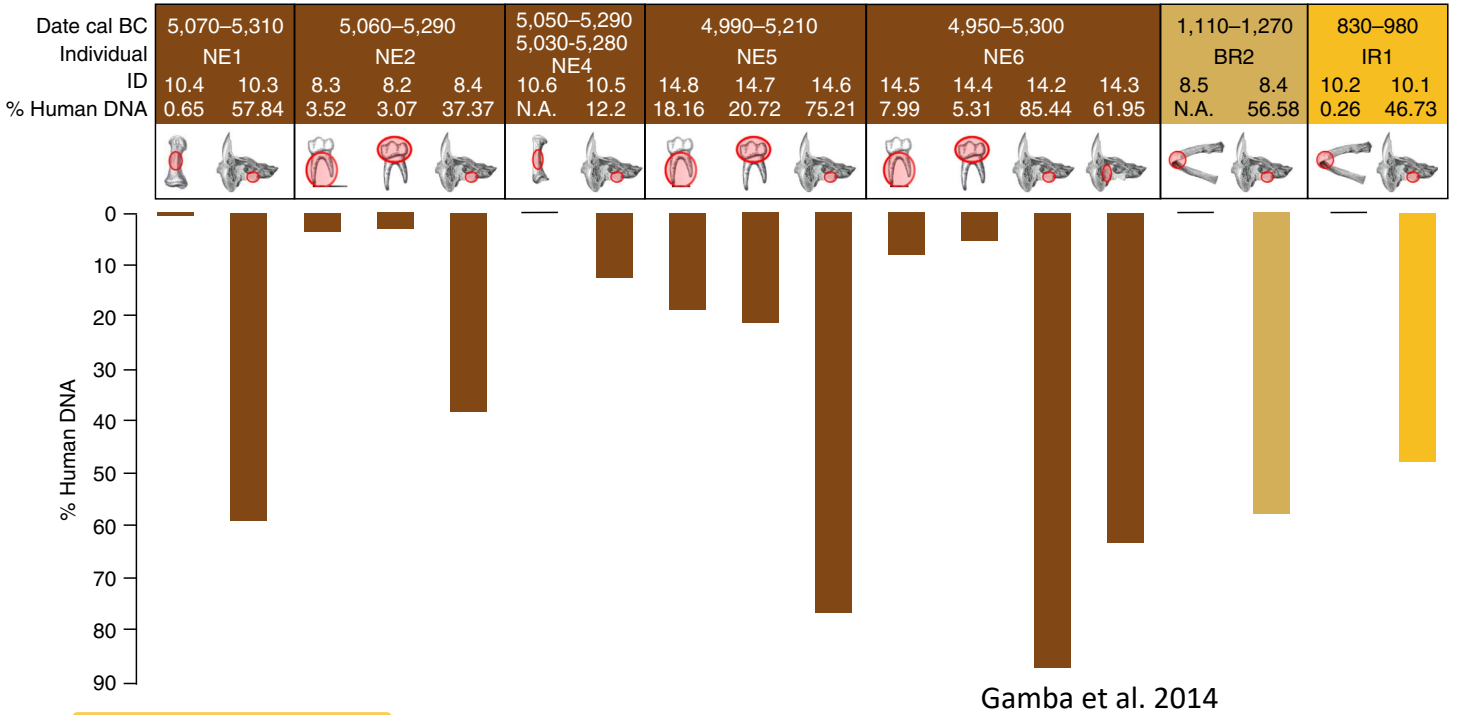
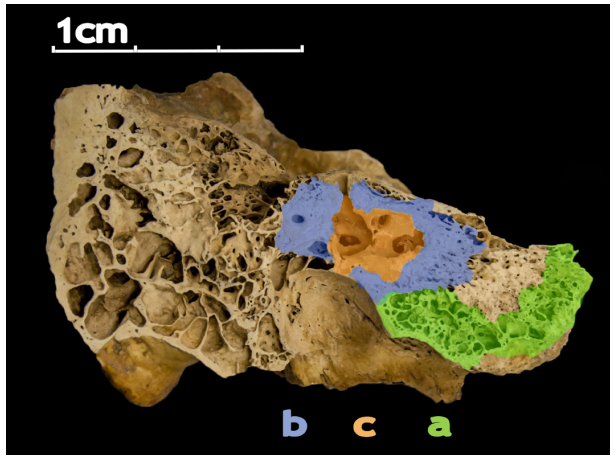


Figure 1 | Petrous bones versus non-petrous bones. Percentage of non-clonal endogenous DNA recovered after shotgun sequencing. The sampled bor tooth portion is circled in red. N.A. indicates that the library did not pass quality assessment for sequencing.

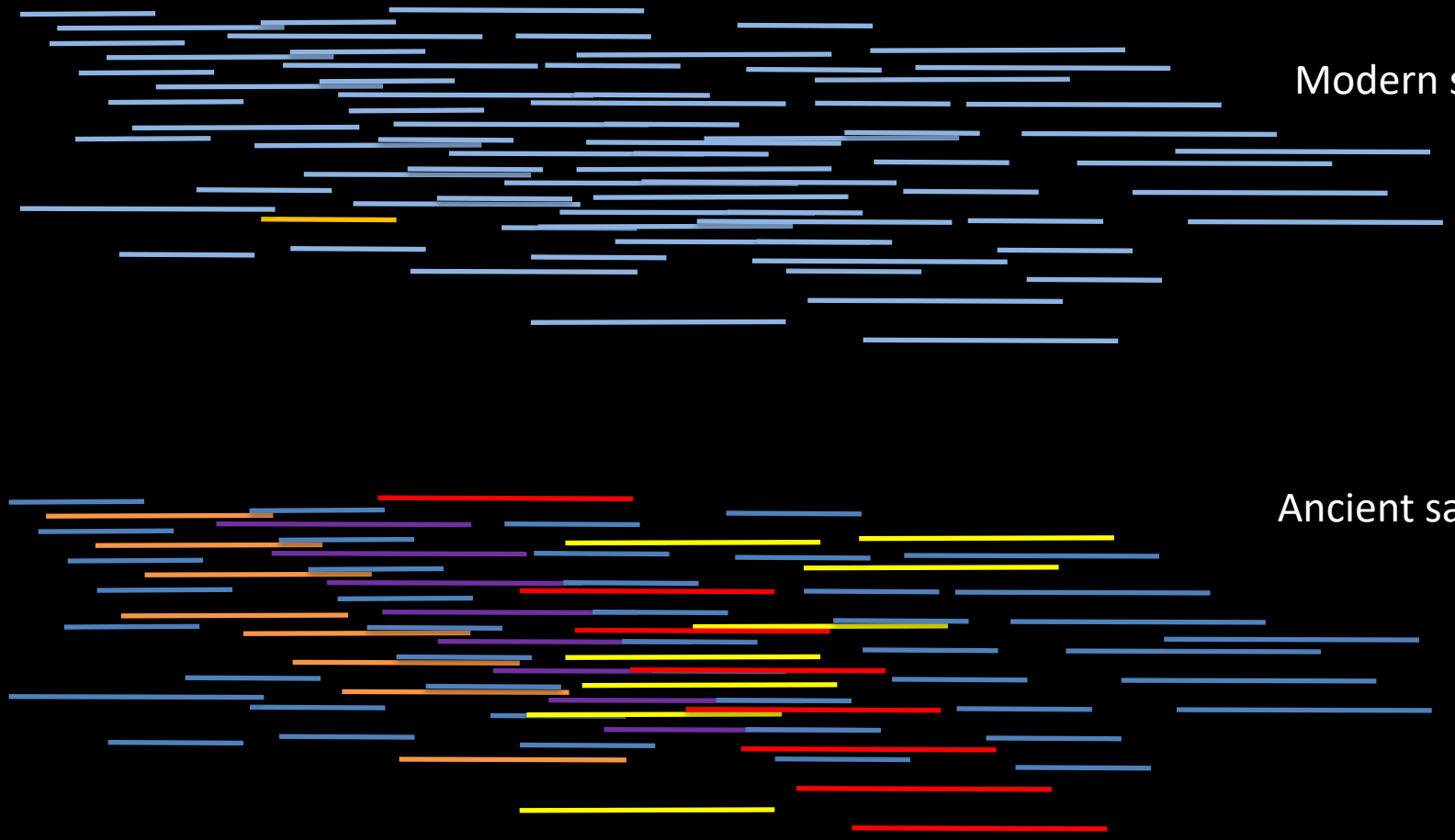


Libraries and complexity

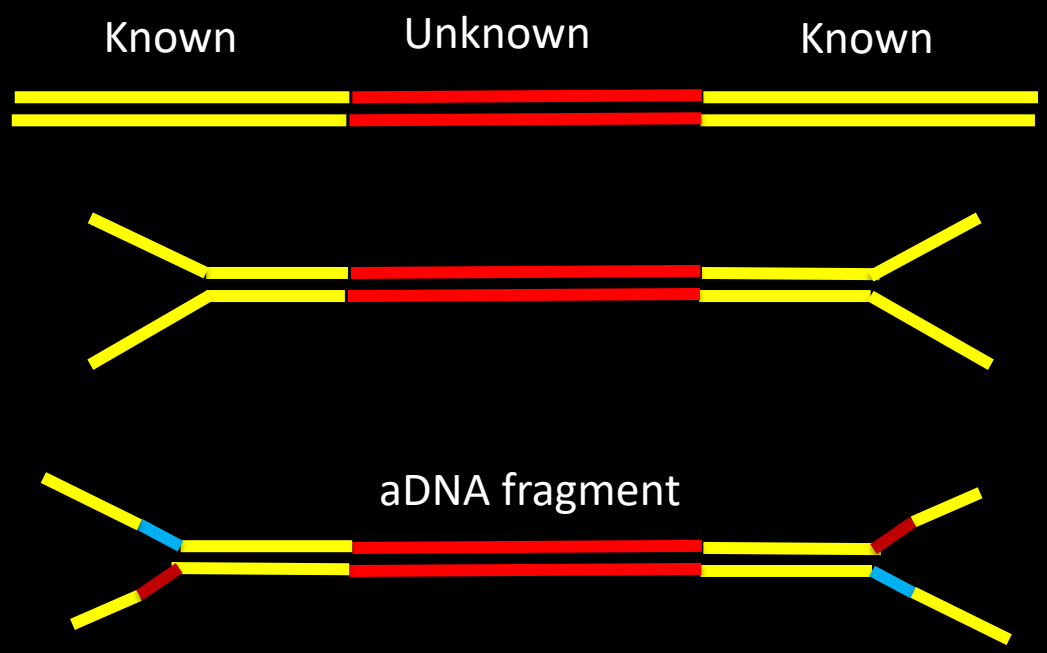
- **Molecular biology library**: A collection of DNA fragments that may be stored... propagated .. .. representing a total swarm of... genome/ mutants/cDNA/.

- **Sequencing library**: A collection of DNA fragments prepared for sequencing, representing (or containing all?) the DNA content of the sample.

Sequencing library



Library fragment between adaptors

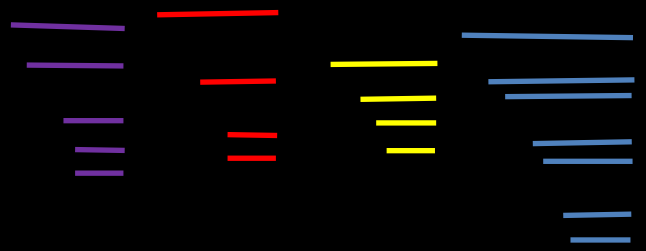


Sequencing adapters
Incl. index

Numbero of unique fragments in library

Sequencing library complexity vs clonality

More library
Same complexity

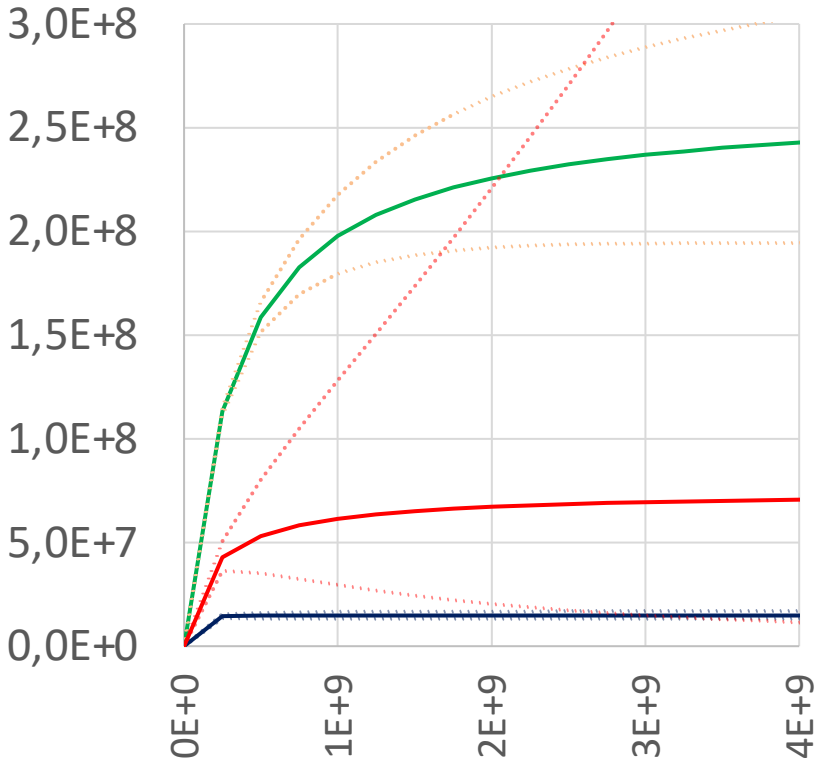


More library,
lower complexity
Higher clonality

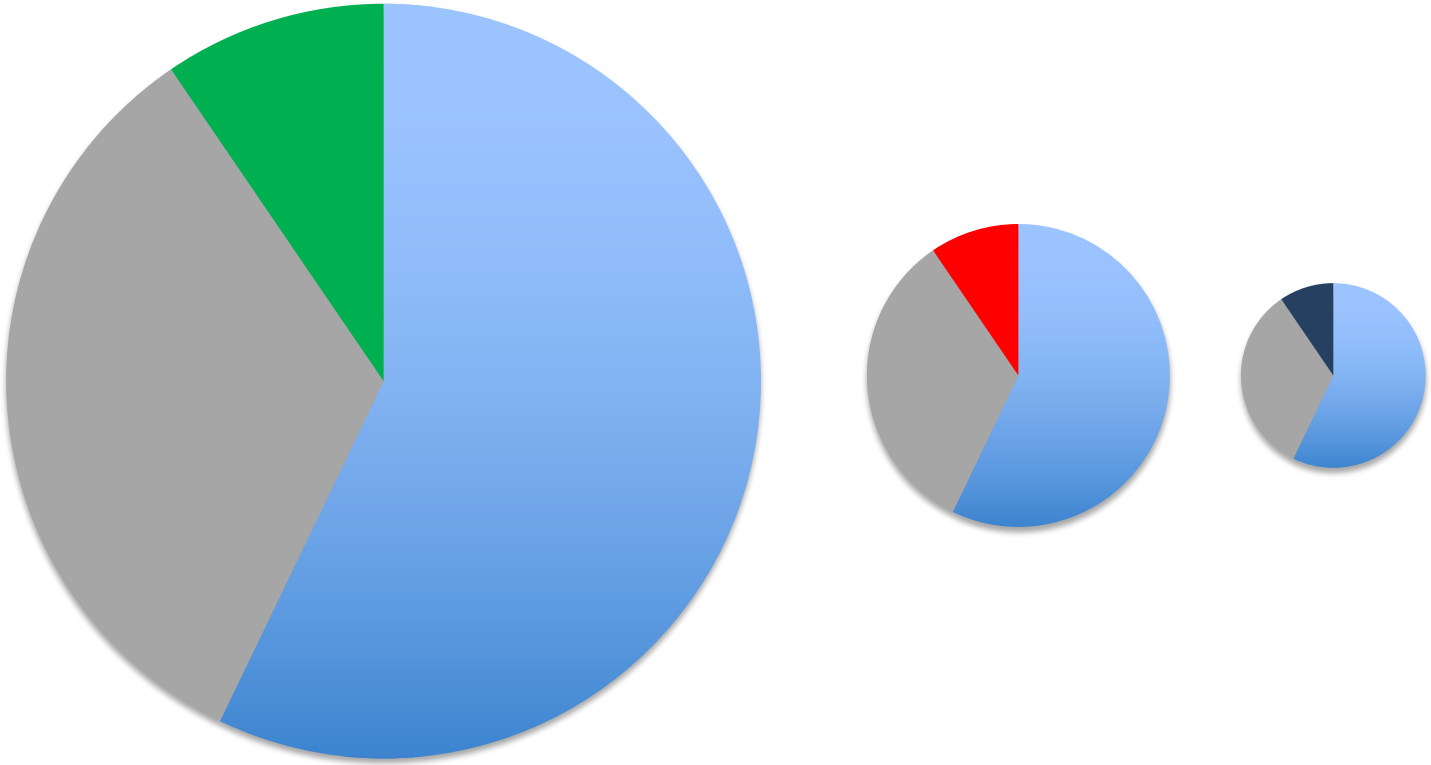
DNA **complexity**

How much data can be squeezed out of the sample?

NEW sequencing reads



Generated sequencing reads



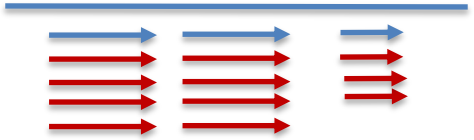
Illumina aDNA sequencing

agtccttggatagtctagtccgatt

Reads/sample (trimmed)	Endo %	Mapping reads	Average length	Coverage (DoC≥1)	DoC	Possible DoC
20,000,000	1	200,000	50 bp	0.3 %	0.003x	1x
20,000,000	15	3,000,000	50 bp	4.5 %	0.05x	10-20x
20,000,000	100	20,000,000	125 bp	76 %	0.8x	>100x

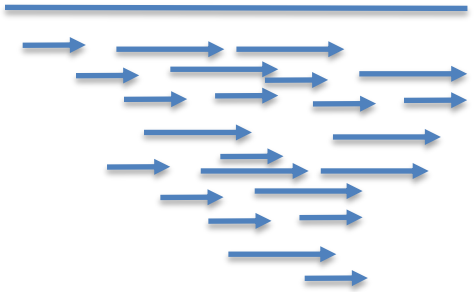
(e.g. 1%)

Low Depth of Coverage (1x)



Clonality

(e.g. 15%)



High Depth of Coverage (10-20x)

Based on slides from Lasse Vinner, head of the sequencing center (on picture)

Summary of aDNA part

- Authentic aDNA is short and damaged
- Sample types and sample quality matter
- Library complexity determines the max seq output
- Risk of contamination
- Clean-lab facilities are required

Standard workflow

- Lab stuff + sequencing
- Call bases and quality scores (FastQ)
- Alignment or *denovo* assembly. Alignment is topic for tomorrow

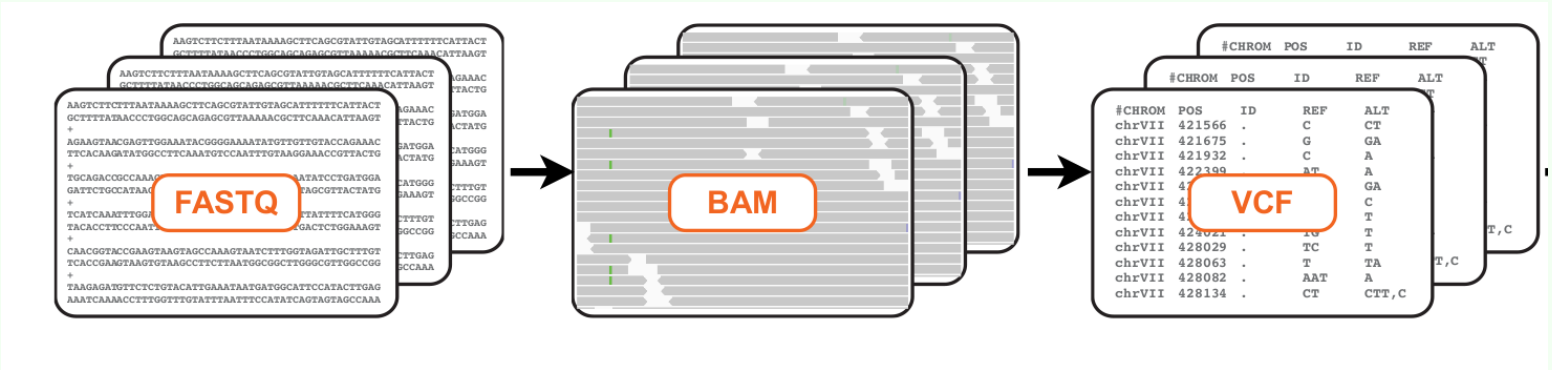
Genotypes

- Generate genotype likelihoods. Uncertainty is topic for friday.
- Estimate allele or haplotype frequencies
- Calculate genotype probabilities
- Call genotypes

Downstream analysis

- Variation detection
- Population genetic analysis

Usual files



- Unmapped reads (`fastq`, fasta files with qualities).
- Mapped reads (`bam`, binary sam files).
- Variant information file (`.vcf` variant call format).

Low level format: Fasta files

```
>MN908947 |Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
CGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA
AGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGC
ATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTC
TTATGTTGGTTGCCATAACAAGTGTGCCTATTGGGTTCCACGTGCTAGCGCTAACATAGG
TTGTAACCATACAGGTGTTGTTGGAGAAGGTTCCGAAGGTCTTAATGACAACCTTCTTGA
AATACTCCAAAAAGAGAAAGTCAACATCAATATTGTTGGTGACTTTAAACTTAATGAAGA
GATCGCCATTATTTTGGCATCTTTTTCTGCTTCCACAAGTGCTTTTGTGGAAACTGTGAA
```

Figure: First sequence of the SARS-CoV-2 virus

- Text file representation of a nucleotide sequence
- Consists of an identifier starting with $>$ followed by actual sequence which might span multiple lines
- A single fasta file might contain multiple entries. E.g chr1 chr2 etc.

Low level format: Fastq files

FASTQ

```
@FC42BF1AAXX:6:1:5:732#0/1              <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN     <-- read (bases)
+
a^_aaaa‘aa‘_aaa_aaa‘__‘‘_‘VBBBBBBBB    <-- quality score
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
ababb_abbbZbabaab^‘aaTaabbaBBBBBBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```

Quality scores/Phred scores

Qscore

- The ASCII values can be interpreted as a probability
- A Q20 (ASCII '5') score is probability of 1%
- The score is the probability, ϵ, that the base is incorrect
- $$Q_{score} = -10log_{10}(\epsilon)$$
- $$\epsilon = 10^{\frac{-Q}{10}}$$

offset of 33 (sometimes 64)

Second round of computer exercises

These exercises revolve around being (still) able to login to the ricco server. And playing around with these lowlevel formats previously described. How long?
`https://github.com/ANGSD/adv_binf_2023_week1`

Alignment

Commonly used aligners

- Bowtie
- BWA
- MAQ
- SOAP2
- Mosaik
- Eland
- DRAGEN

Alignment

Commonly used aligners

- Bowtie
- BWA
- SOAP2
- DRAGEN

Burrows-Wheeler Alignment: are extremely fast. but not commonly used ones are not quality sensitive!

Alignment file

Alignments of the reads to a reference genome

an alignment file includes

reads	`TTTGTTCTTTCTTTCTCTCTAGTCTTCTT ...`
Qscore	`NVFVN]^]‘^_]^^U]]‘][_VS[_^Z]_ ...`
start position	chr4 53351385
multiple best hits	1
Number of mismatch	2
sequence strand	-
mapping quality	V

SAM/BAM format 1/3

```
@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
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_topha
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    *
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCA
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDE
    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:553527
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGG
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHH
    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    *
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACA
C    DDDDDDDDDCCDDDDDDDDDEEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFG
    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       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAAT
accepted_hits.sam
```

Figure: Example of a small sam file. Header part starts with @. The actual alignment information for the reads follows after the header. One read correspond to one line

SAM/BAM format 2/3

Tag	Description
@HD	The header line. The first line if present.
VN*	Format version. *Accepted format*: /^[0-9]+\.[0-9]+$/.
SO	Sorting order of alignments. *Valid values*: unknown (default), unsorte coordinate. For coordinate sort, the major sort key is the RNAME field, by the order of @SQ lines in the header. The minor sort key is the POS fi with equal RNAME and POS, order is arbitrary. All alignments with '*' in alignments with some other value but otherwise are in arbitrary order.
GO	Grouping of alignments, indicating that similar alignment records are group file is not necessarily sorted overall. *Valid values*: none (default), query (alig by QNAME), and reference (alignments are grouped by RNAME/POS).
@SQ	Reference sequence dictionary. The order of @SQ lines defines the alignment
SN*	Reference sequence name. Each @SQ line must have a unique SN tag. The valu in the alignment records in RNAME and RNEXT fields. Regular expression:
LN*	Reference sequence length. *Range*: [1,2^{31}-1]
AS	Genome assembly identifier.
M5	MD5 checksum of the sequence in the uppercase, excluding spaces but inclu
SP	Species.
UR	URI of the sequence. This value may start with one of the standard protoco If it does not start with one of these protocols, it is assumed to be a file-sys
@RG	Read group. Unordered multiple @RG lines are allowed.
ID*	Read group identifier. Each @RG line must have a unique ID. The value of I tags of alignment records. Must be unique among all read groups in header s IDs may be modified when merging SAM files in order to handle collisions.
CN	Name of sequencing center producing the read.
DS	Description.
DT	Date the run was produced (ISO8601 date or date/time).
FO	Flow order. The array of nucleotide bases that correspond to the nucleo flow of each read. Multi-base flows are encoded in IUPAC format, and non- various other characters. *Format*: /*|[ACMGRSVTWYHKDBN]+/
KS	The array of nucleotide bases that correspond to the key sequence of each r
LB	Library.
PG	Programs used for processing the read group.
PI	Predicted median insert size.
PL	Platform/technology used to produce the reads. *Valid values*: CAPILLARY SOLID, HELICOS, IONTORRENT, ONT, and PACBIO.
PM	Platform model. Free-form text providing further details of the platform/te
PU	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). U
SM	Sample. Use pool name where a pool is being sequenced.
@PG	Program.
ID*	Program record identifier. Each @PG line must have a unique ID. The value alignment PG tag and PP tags of other @PG lines. PG IDs may be modified files in order to handle collisions.

SAM/BAM format 3/3

SAM format

QNAME	ID of the read
FLAGS	flag of the alignment
RNAME	Reference name
POS	1-indexed position of leftmost first matching position in sequence
MAPQ	mapping quality (phred scaled)
CIGAR	Alignment info
RNEXT	mate reference name (paired end)
RPOS	mate position (paired end)
ISIZE	inferred insert size
SEQ	the sequence
QUAL	qscore of the sequence

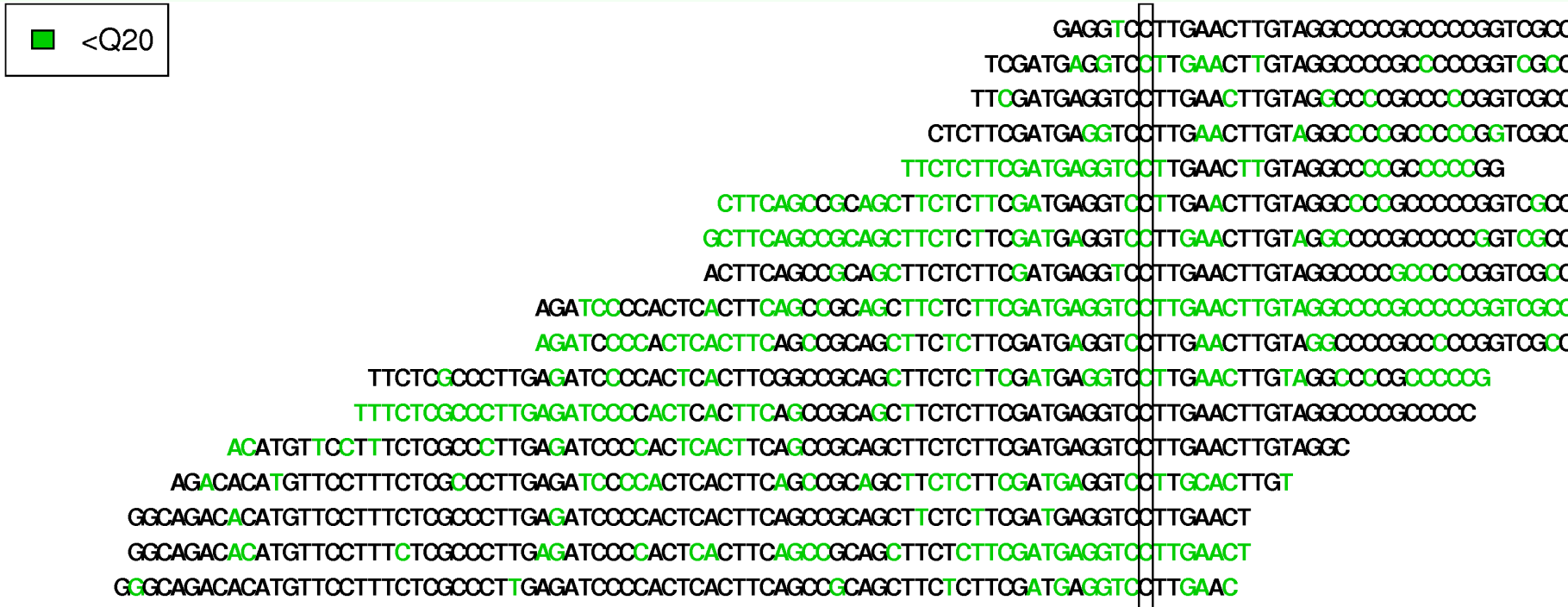
Mapped reads

My definitions (The literature is not consistent)

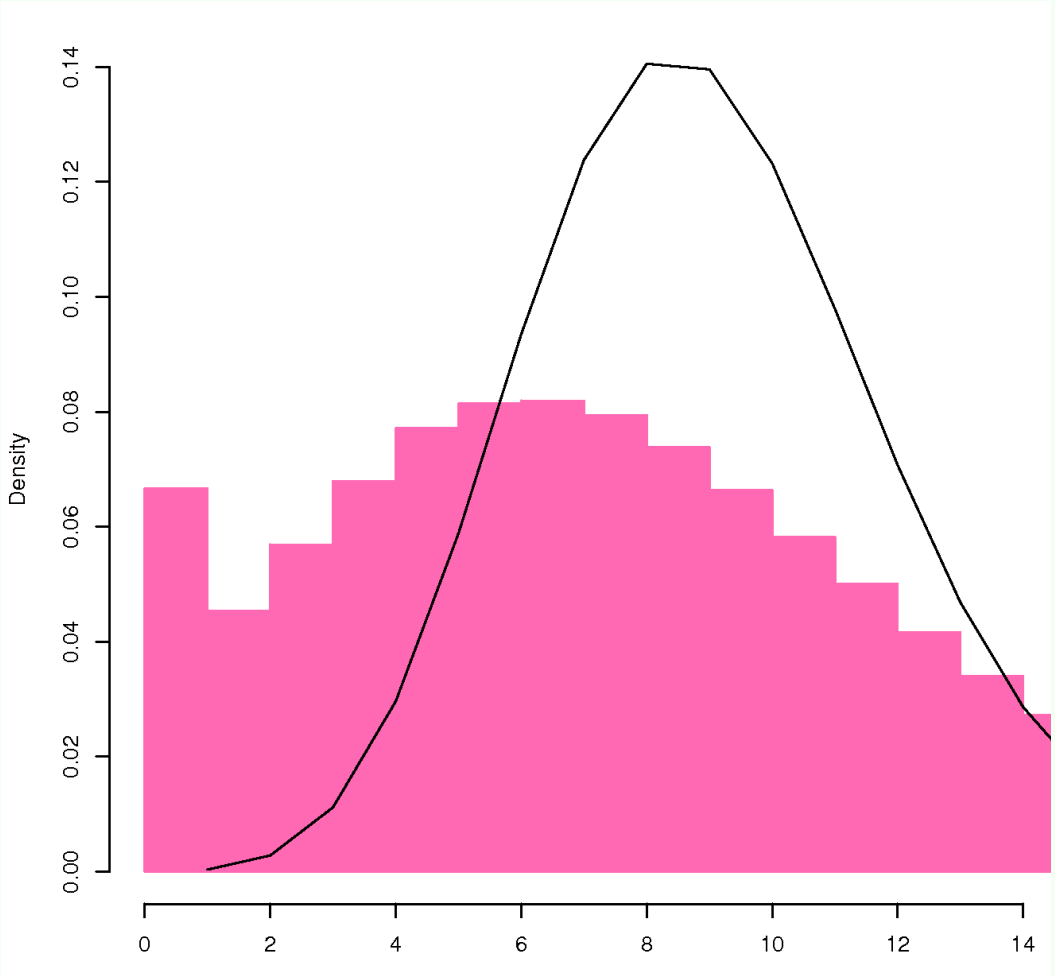
Depth The number of reads that maps to a position

Counts The number of different alleles mapped to a position

Coverage The fraction of the genome (region) with data



Distribution of depths



- NGS data with mean depth 8
- If the reads are random then we expect the depth to follow a Poisson distribution
- you need a very large depth to cover all sites

Figure: black line:expected. pink: observed

Definitions

* An *allele* is a variant form of a gene.
* A *genotype* are the alleles for a specific individual.
* *genotype calling* is the process of determining the underlaying 'true' genotype that has generated the data we observe.

SNP-chip and high throughput sequencing

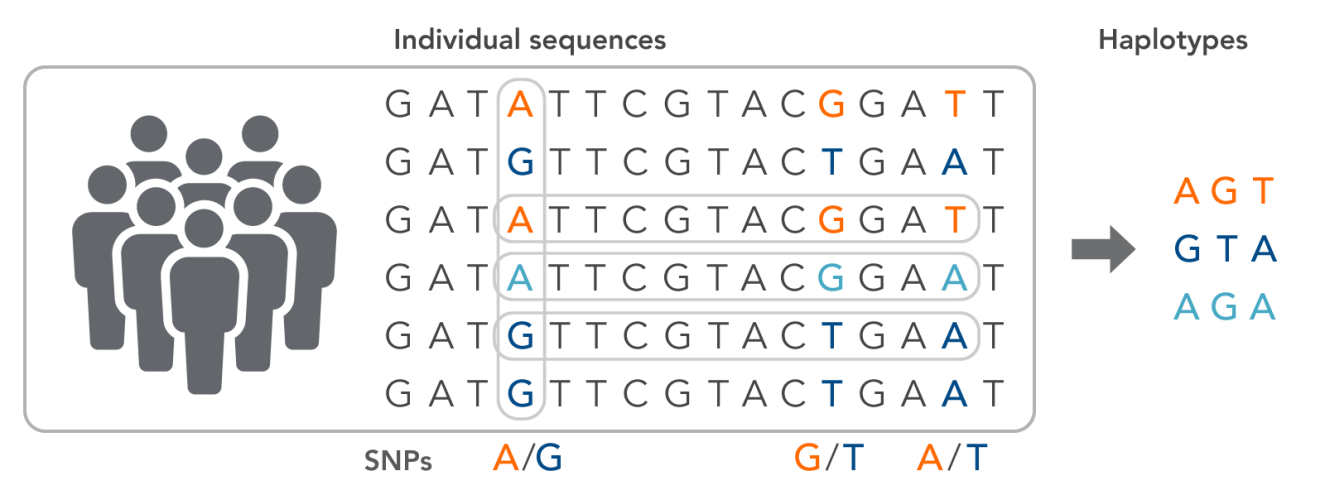
SNP-chip Genotypes are readily available for a predetermined list of of positions.

High-throughput sequencing Data are short sequences that are sampled from either chromosome with replacement.

Question

Questions?

- Identify some of the alleles from the figure?
- Identify some of the genotypes.
- Calculate allele frequencies for the variable positions.



VCF format

(a) VCF example

Header:

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##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">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Header + Body:

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

(b) SNP

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

POS	REF	ALT
2	C	T

(c) Insertion

```
12345
AC-GT
ACTGT
  ^
```

POS	REF	ALT
2	C	CT

(d) Deletion

```
1234
ACGT
A--T
 ^^
```

POS	REF	ALT
1	ACG	A

(e) Replacement

```
1234
ACGT
A-TT
 ^^
```

POS	REF	ALT
1	ACG	AT

(f) Large structural variant

Alignment

```
   100       110       120       290       300
    .         .         .         .         .
ACGTACGTACGTACGTACGTACGT[...]ACGTACGTACGTAC
ACGT------------------------[...]----------GTAC
```

VCF representation

POS	REF	ALT	INFO
100	T	<DEL>	SVTYPE=DEL;END=299

(g) Resolving ambiguity

Alignment

```
1234567890
TTTCCCTCTA
CTTACCT--A
^  ^   ^^
```

Possible representation

POS	REF	ALT
1	TTTCCCTCT	CTTACCTA

Possible representation

POS	REF	ALT
1	T	C
4	C	A
7	TCT	T

Recommended VCF representation

POS	REF	ALT
1	T	C
4	C	A
5	CCT	C

Third and last round of computer exercises

These exercises revolve around alignment files and variant information files previously described. How long?
`https://github.com/ANGSD/adv_binf_2023_week1`