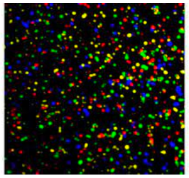
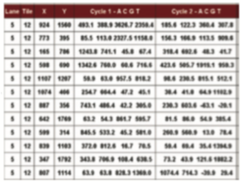
.bcl file

binary base call format

Raw output of sequencing machine; contains the **base call** and the **quality (confidence)** of the base call after each cycle (for example 2 x 100 chemistry has 100 forward and 100 reverse cycles) by cluster. Can be hundreds of GB worth of data.





Lane	Tile	X	Y	Cycle 1 - A C G T	Cycle 2 - A C G T
5	12	924	1560	493.1 388.9 3626.7 2350.4	185.6 122.3 360.4 307.8
5	12	773	395	85.5 113.0 2327.5 1158.0	156.3 166.9 113.5 909.6
5	12	165	786	1243.8 741.1 45.8 67.4	318.4 692.6 48.3 41.7
5	12	598	690	1342.6 760.0 60.6 716.6	423.6 505.7 1919.1 959.3
5	12	1107	1207	58.9 63.0 957.5 818.2	98.6 230.5 815.1 512.1
5	12	1074	466	254.7 664.4 47.2 45.1	38.4 41.8 64.9 1102.9
5	12	887	356	743.1 486.4 42.2 305.0	230.3 603.6 -43.1 -20.1
5	12	642	1769	63.2 54.3 861.7 595.7	81.5 86.0 54.9 385.4
5	12	599	314	845.5 533.2 45.2 581.0	260.9 560.9 13.0 78.4
5	12	839	1103	372.0 812.6 16.7 70.5	58.4 69.4 35.4 1394.9
5	12	347	1792	343.8 706.9 108.4 638.5	73.2 43.9 121.6 1882.2
5	12	807	1114	63.9 63.8 828.3 1369.0	1074.4 714.3 -39.9 29.4

Demultiplexing is the process of sorting base calls into separate files by their unique indices

Demultiplexing

Performed with Illumina's bcl2fastq tool. Almost always done by the sequencing centre.

.fastq file

also shortened as .fq

Information contained in .bcl file organized by sample and by sequence. Contains **every sequence read** (A,T,C, G, and N) and **the quality score (confidence) of each base call**. Quality scores represent the probability that the base was called in error. The quality score is reported in ASCII characters, 42 single digit numbers, characters, and letters. This makes fastq format easily human-readable, because each base (A, T, C, G, or N) is associated with a single digit quality score (!, *, 6, C, J, etc).

In fastq format, each sequence is represented by four lines:

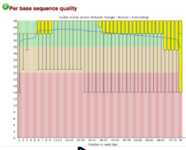
1. A header with seq info, including cluster coordinates from flow cell
2. The sequence
3. Qual score separater "+"
4. The quality scores

Example of one read in .fastq format:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

Single end (SE) sequencing runs produce one .fastq file (**R1**) per read, and **paired end (PE)** runs produce two files (**R1 - forward reads** and **R2 - reverse reads**)

Sequencing reads will be of various quality and need to be assessed. The program **fastqc** is used to visualize different sequence quality metrics



Quality control

Based on the quality metrics and the type of data you have (PE, SE, ancient, or modern), you will **trim the reads to remove adapter sequences, N base calls, and reads with low base quality** (low confidence in the call). If you have ancient PE data, you may also want to merge the reads where they overlap.

quality filtered .fastq file

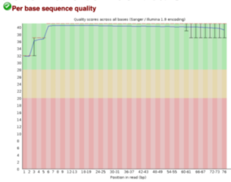
also shortened as .fq

Common quality issues in aDNA that should be trimmed and filtered out of your fastq files:

-Adapter readthrough = your sequencing chemistry is longer than your fragments so the adapter sequence is read as part of the actual sequence.

-Ns on the ends of reads: complete adapter readthrough and poor quality data can result in Ns, meaning the sequencer could not call any base. These are not bases with typical aDNA "damage"

-Low quality base calls: low quality DNA, poor library construction, insufficient purificaiton can result in the sequencing machine not being able to make confident calls.



You may want to **merge** your forward and reverse reads. aDNA is short and reads are more likely to overlap. This helps remove contaminant DNA, which is generally longer.

Now that you are confident in your sequence read quality, you want to **map the reads to the reference genome of your target organism**.

There are different mapping algorithms, but essentially matching and mismatching bases between ref and seq reads are tabulated for an overall alignment score for each read. Common tools include BWA and bowtie.

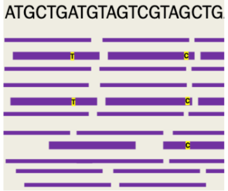
Mapping

A reference genome is a high quality genome assembly. There is generally one reference genome per organism that all scientists use. For example, there is one human reference genome that is regularly being updated as more sections of the genome are annotated (hg38).

.sam/.bam files

and their indexes (.sai and .bai)

SAMs are tsv (tab-separated values) **files that contain a header and the alignment score for every read. A BAM file is the binary (compressed) version of a SAM file**. SAMs can be converted to BAMs, and vice versa. BAM files are not human readable (if you open it up it is a bunch of nonsense). Alignment files can be huge and take a lot of memory to parse. Compression into a BAM makes computation faster and more efficient. SAM/BAMs have associated index files (.sai/bai) used for increased computatiion efficiency.



SAM format includes header info beginnging with the "@" symbol; each row corresponds to a read, and each column stores information about the read alignment.

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
```

Unfiltered SAM/BAM files contain a lot of information we don't need (unmapped reads) and a lot of poor quality alignments (reads with bad alignments scores) that we don't want to include in final analyses.

Common to filter out:

-Unmapped reads

-Reads with low mapping quality scores: low confidence that the read actually goes where placed.

Filtering bams

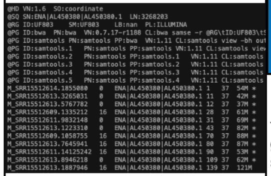
-Duplicate reads: PCR and optical duplicates can make your data seem higher coverage. Generally we only want unique reads in our analyses.

-Reads with more than one alignment: one read can be aligned equally well to multiple segments of the genome. Since we can't be sure where they belong, they are often removed.

quality filtered .sam/.bam files

and their indexes (.sai and .bai)

You now have a quality filtered (QF) alignment file (.sam/.bam)!



```
@HD VN:1.6  SO:coordinate
@SQ SN:ENA|AL450380|AL450380.1  LN:3268203
@RG ID:UF803    SM:UF803    LB:nan  PL:ILLUMINA
@PG ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa samse -r @RG\tID:UF803\tS
@PG ID:samtools PN:samtools PP:bwa  VN:1.11 CL:samtools view -bh out
@PG ID:samtools.1   PN:samtools PP:samtools VN:1.11 CL:samtools view
@PG ID:samtools.2   PN:samtools PP:samtools.1   VN:1.11 CL:samtools
@PG ID:samtools.3   PN:samtools PP:samtools.2   VN:1.11 CL:samtools
@PG ID:samtools.4   PN:samtools PP:samtools.3   VN:1.11 CL:samtools
@PG ID:samtools.5   PN:samtools PP:samtools.4   VN:1.11 CL:samtools
M_SRR15512614.1855080   0   ENA|AL450380|AL450380.1 1   37  54M *
M_SRR15512613.3265031   0   ENA|AL450380|AL450380.1 11  37  42M *
M_SRR15512613.5767782   0   ENA|AL450380|AL450380.1 12  37  37M *
M_SRR15512609.1335212   16  ENA|AL450380|AL450380.1 28  37  61M *
M_SRR15512611.9832148   0   ENA|AL450380|AL450380.1 31  37  69M *
M_SRR15512613.1223310   0   ENA|AL450380|AL450380.1 43  37  82M *
M_SRR15512609.1058755   16  ENA|AL450380|AL450380.1 70  37  88M *
M_SRR15512613.7645941   16  ENA|AL450380|AL450380.1 80  37  87M *
M_SRR15512611.14125242  16  ENA|AL450380|AL450380.1 90  37  53M *
M_SRR15512613.8946218   0   ENA|AL450380|AL450380.1 109 37  62M *
M_SRR15512613.1887946   16  ENA|AL450380|AL450380.1 139 37  121M
```

samtools is the primary tool for sam/bam file manipulation and is used for all NGS data, modern and ancient.

A cleaned .sam/.bam file is generally analysis-ready, meaning that now you can calculate mapping statistics (% reference coverage, depth of coverage, read length distribution, etc) and begin genome analysis of point mutations (SNPs aka SNVs) or structural vairants.

Ancient DNA that has not been fully treated with UDG to remove C->T and G->A transitions caused deamination usually requires one more step to rescale these positions so they are not called as variants.

Rescaling with MapDamage

Following .sam/.bam filtering, it is typical to generate mapping statistics using a tool called Qualimap.

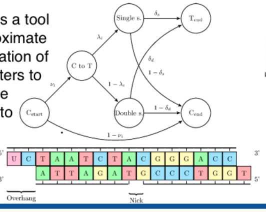
The aDNA bioinformatic workflow is pretty standardized until this point. Contamination assessment, however, is organism-dependent and can be done many ways.

Decontamination with pmdtools

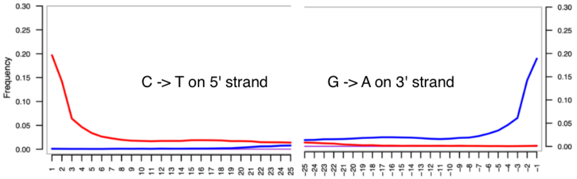
rescaled quality filtered .bam files and their indexes (.bai)

C->T and G->A transitions accumulate near the ends of aDNA fragments. Miscoding lesion may not be flagged as poor quality from sequencing or mapping, so they remainin in the quality filtered reads.

MapDamage2 is a tool that uses approximate bayesian estimation of damage paramters to identify probable transitions due to damage.



MapDamage2 modifies the .bam file such that **base quality scores** (data from the sequencing run that come from the fastq files that are also stored in the resulting .sam/.bams) **are reduced based on the probability that they resulted from a miscoding lesion.**



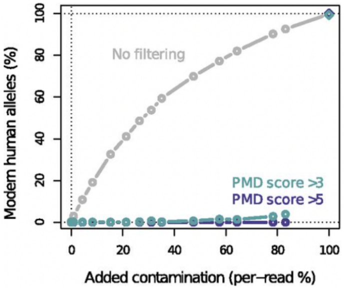
MapDamage2 also produces damage plots, showing the frequency of miscoding lesions on the read ends.

decontaminated .bam files and their indexes (.sai/.bai)

Miscoding lesions due to degredation can also be used directly for read authentication.

PMDtools (post-mortem damage tools) removes contaminant reads from .sam/.bam files using a likelihood framework that models three processes:

1. C -> T changes resulting from post mortem damage
2. True C -> T biological polymorphisms (SNPs/SNVs)
3. C -> T changes caused by sequencing errors



PMDtools assigns each read a PMD score, and reads with a score below the threshold are discarded.

This approach has been shown to reduce contamination to aDNA datasets to negligible amounts while maintain important biological data that could be lost under more conseravtive approaches.

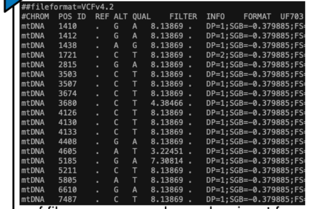
PMDtools modifies the .sam/.bam by removing low-scoring reads.

Variant calling

.vcf files

Variant call format (VCF) files are tsv files with a standardized format used to store information about variants observed in reads after they have been aligned to a reference genome. VCF files can store information about point mutations (SNPs/SNVs), insertions, deletions, and large structural variants.

Like .sam/.bam files, .vcf files have headers with mandatory and optional information. After the header, each row corresponds to a variant and provides information on the genomic position of the variant, the reference allele, and data used to assess the quality and confidence of the variant call, such as read depth and genotype likelihood. There are MANY variant calling tools. Some are better for aDNA.



```
##fileformat=VCFv4.2
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  UF703
mtDNA   1410    .   G   A   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   1412    .   G   A   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   1438    .   A   G   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   1721    .   C   T   8.13869 .   DP=2;SGB=-0.379885;FS
mtDNA   2815    .   G   A   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   3503    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   3507    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   3674    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   3680    .   C   T   4.38466 .   DP=1;SGB=-0.379885;FS
mtDNA   4126    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   4130    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   4133    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   4408    .   G   A   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   4605    .   A   T   3.22451 .   DP=1;SGB=-0.379885;FS
mtDNA   5185    .   G   A   7.30814 .   DP=1;SGB=-0.379885;FS
mtDNA   5211    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   5805    .   A   T   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   6610    .   G   A   8.13869 .   DP=1;SGB=-0.379885;FS
mtDNA   7487    .   C   T   8.13869 .   DP=1;SGB=-0.379885;FS
```

.vcf files are commonly used as input for analytical tools, such as HaploGrep2, PLINK, and SNPeff.