Bioinformatic challenges associated with ancient DNA

Computational Challenges Workshop
9 November 2015

Karola Kirsanow
Christian Sell
AG Palaeogenetik

JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Computational challenges unique to aDNA

- The nature of the question: ascertainment bias and other biases affecting our priors
- The nature of the molecule: biases in preservation and endogeneity unique to aDNA
- The nature of the analysis: biases introduced during NGS and downstream bioinformatics analyses

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Computational challenges unique to aDNA

- The nature of the question: ascertainment bias and other biases affecting our priors
- The nature of the molecule: biases in preservation and endogeneity unique to aDNA
- The nature of the analysis: biases introduced during NGS and downstream bioinformatics analyses

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Ancient DNA

Modern DNA

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Computational challenges unique to aDNA

- The nature of the question: ascertainment bias and other biases affecting our priors
- The nature of the molecule: biases in preservation and endogeneity unique to aDNA
- The nature of the analysis: biases introduced during NGS and downstream bioinformatics analyses

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Addressing computational challenges unique to aDNA

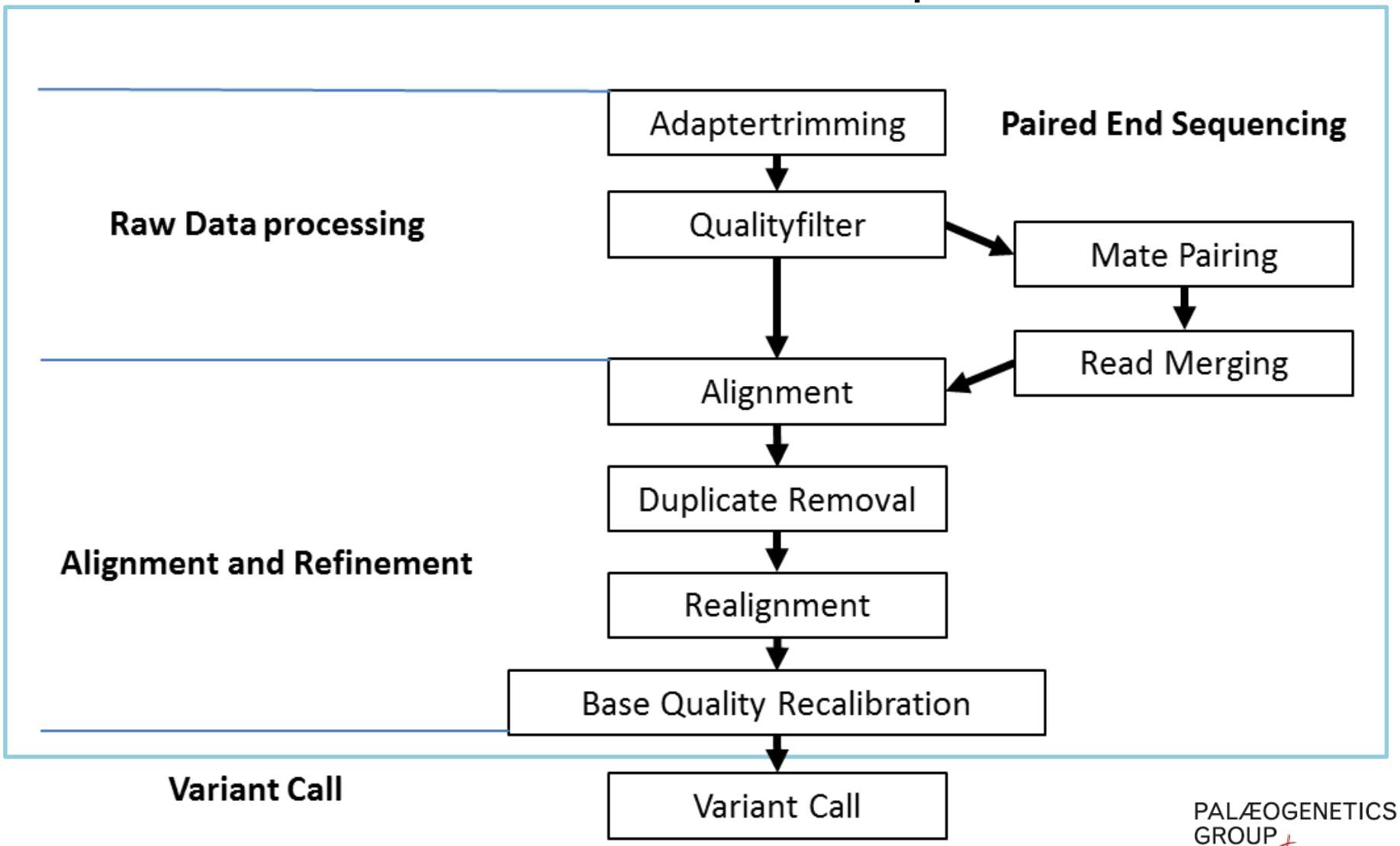
- The nature of the question: formulate research questions incorporating archaeological and palaeo-population genetic data

- The nature of the molecule: characterize and correct for stereotypical damage patterns

- The nature of the analysis: develop appropriate reference genomic data sets and devise calling and recalibration methods suited to ancient DNA

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

The Bioinformatic Pipeline



PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

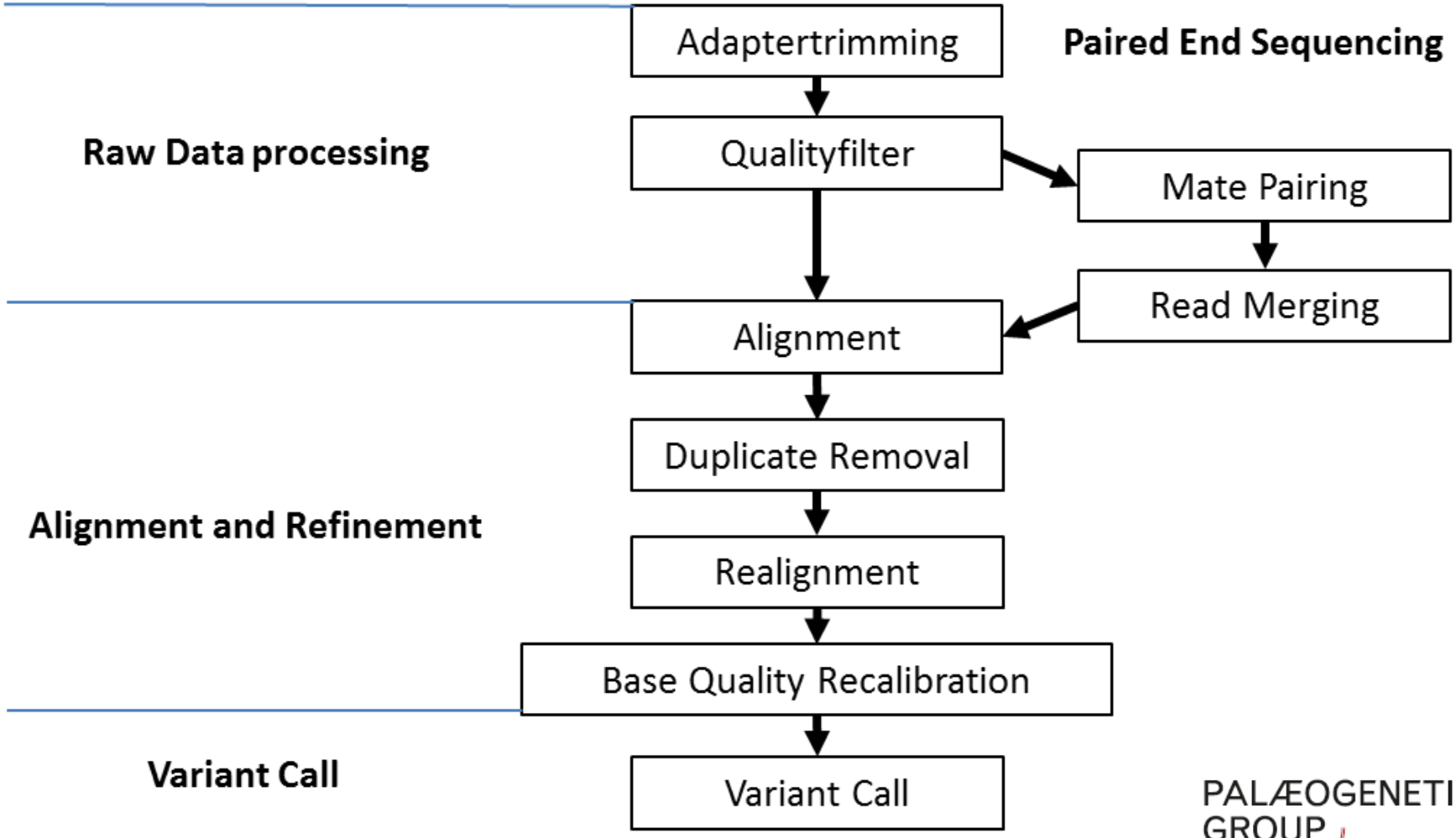
Capturing ancient patterns of genomic variation

= truth

= filtered

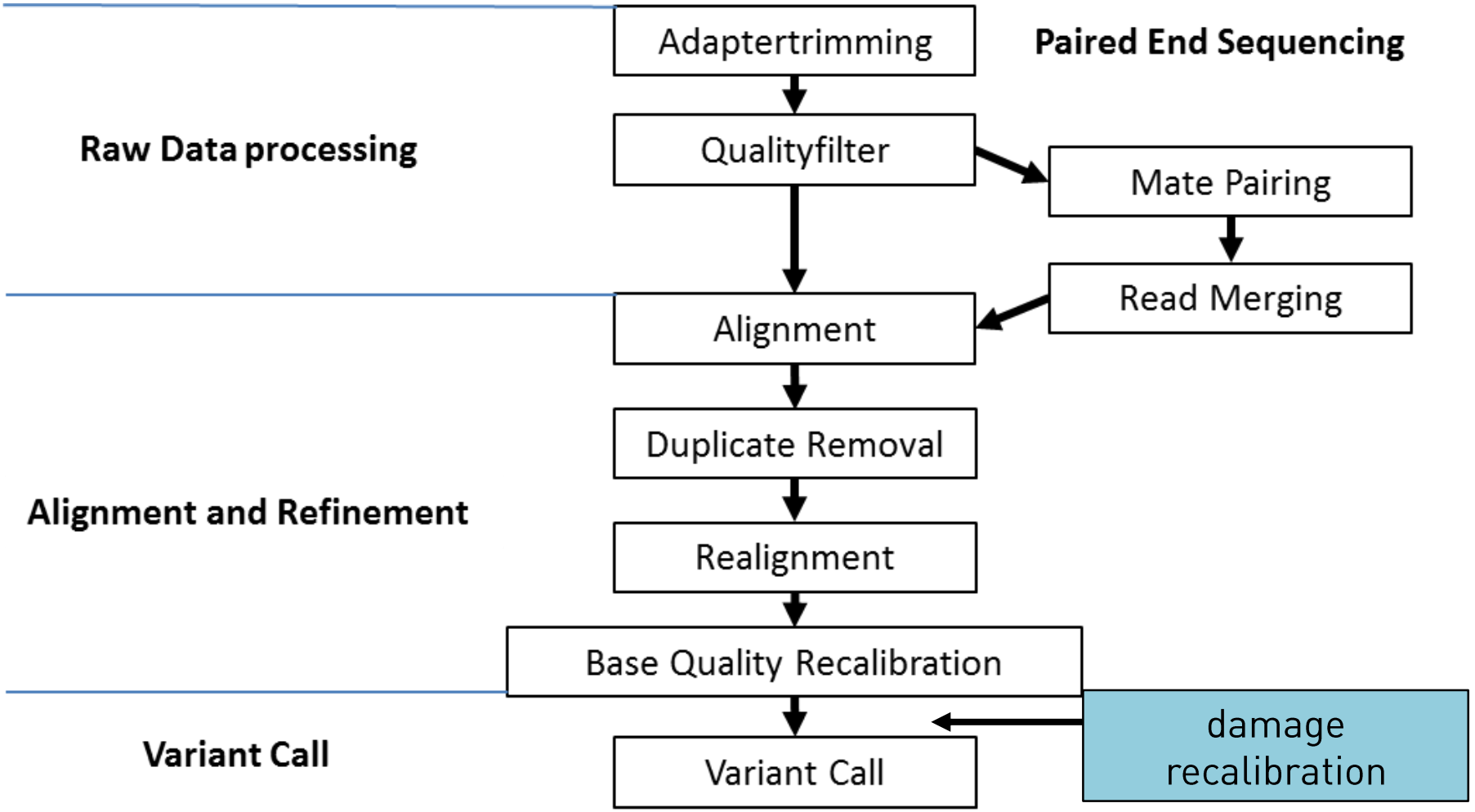
PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

The Bioinformatic Pipeline



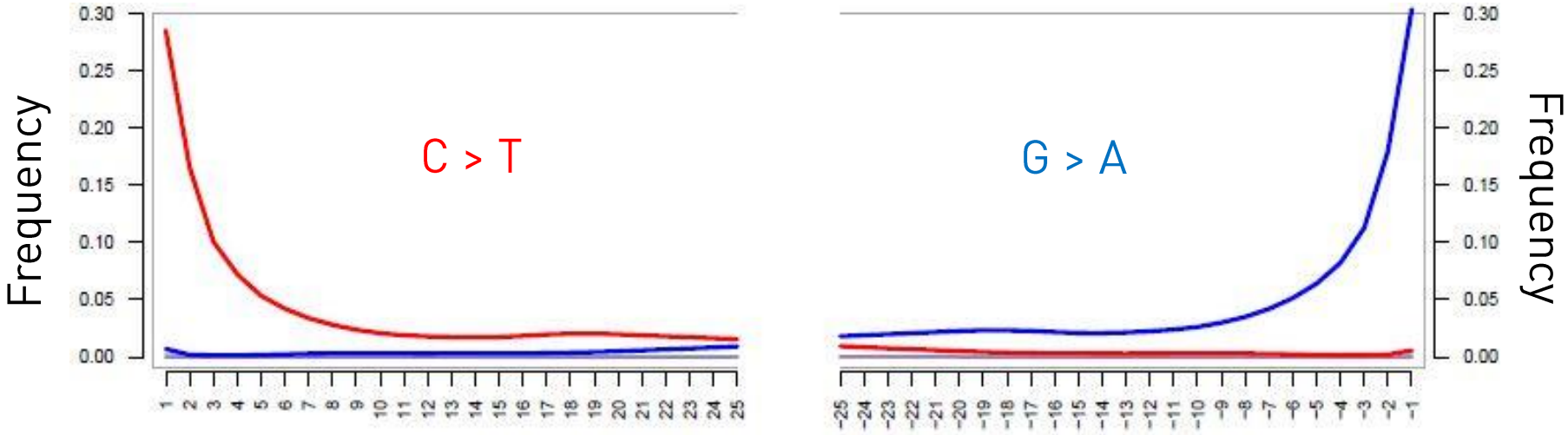
PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

The Bioinformatic Pipeline



PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

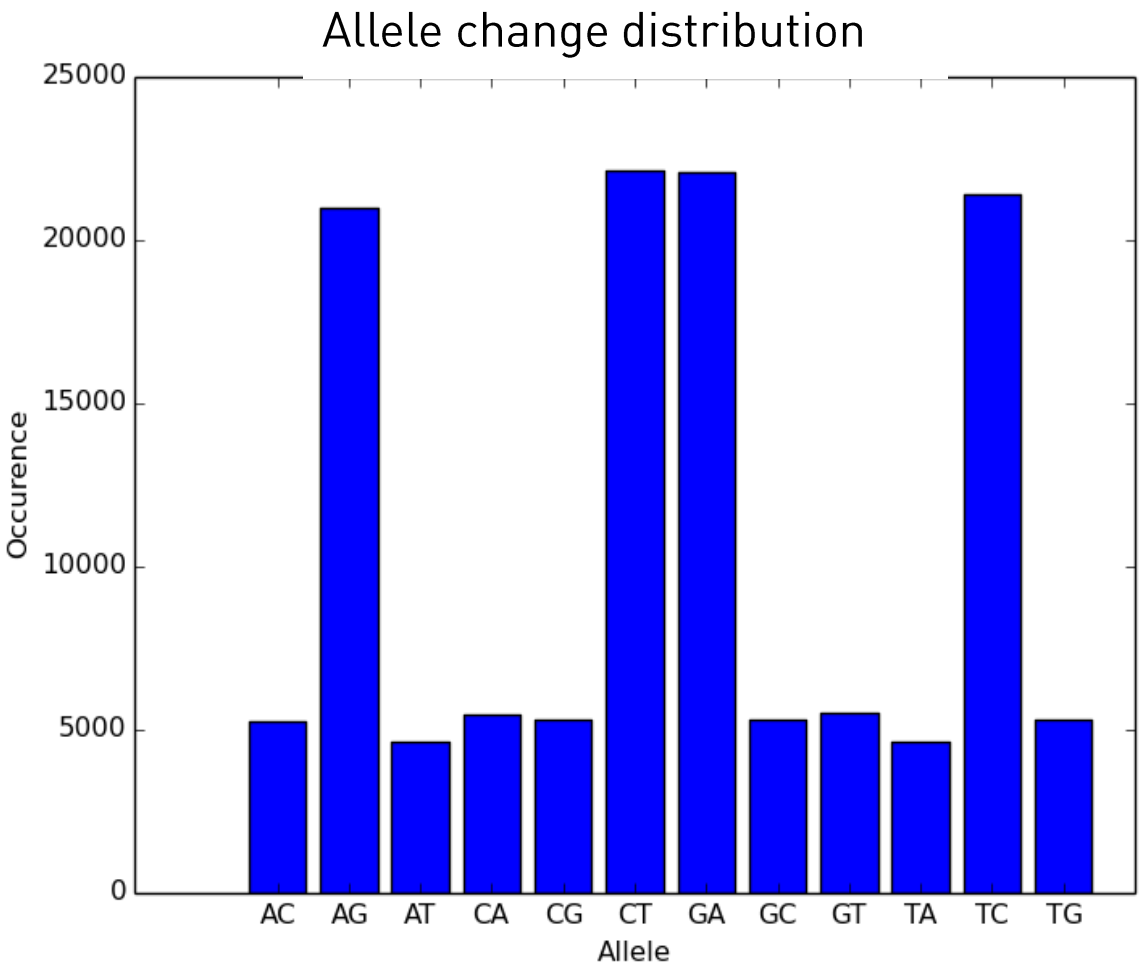
Damage recalibration



Adjust base quality scores in C>T and G>A transitions according to damage patterns, following an empirical model

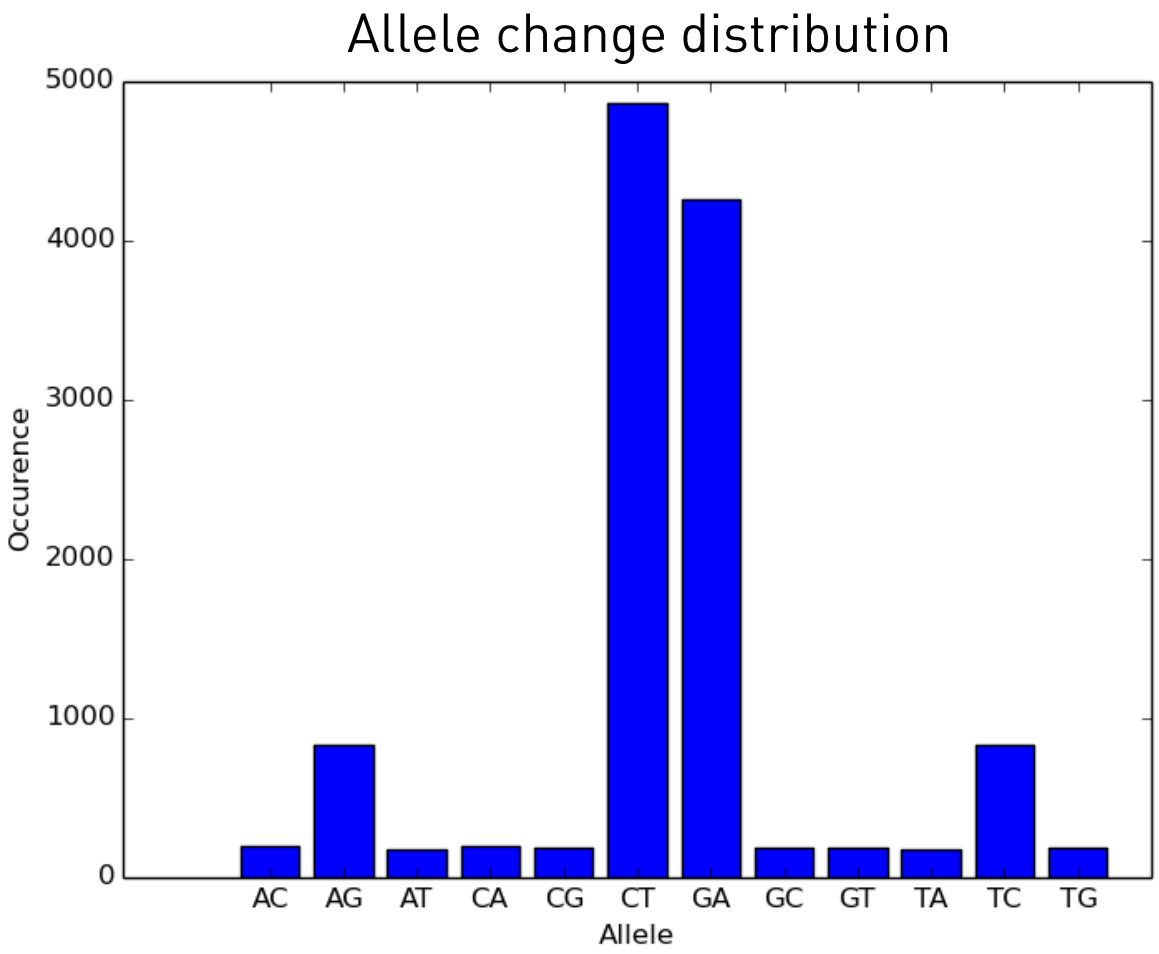
PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Modern sample NA18534



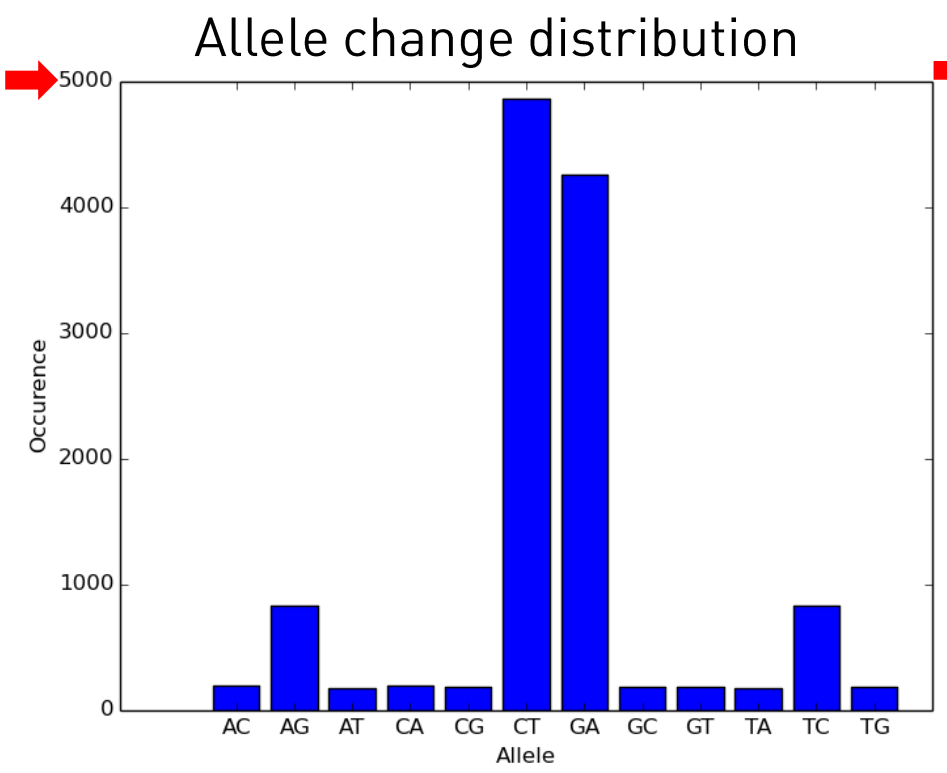
PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

aDNA sample Klei 10

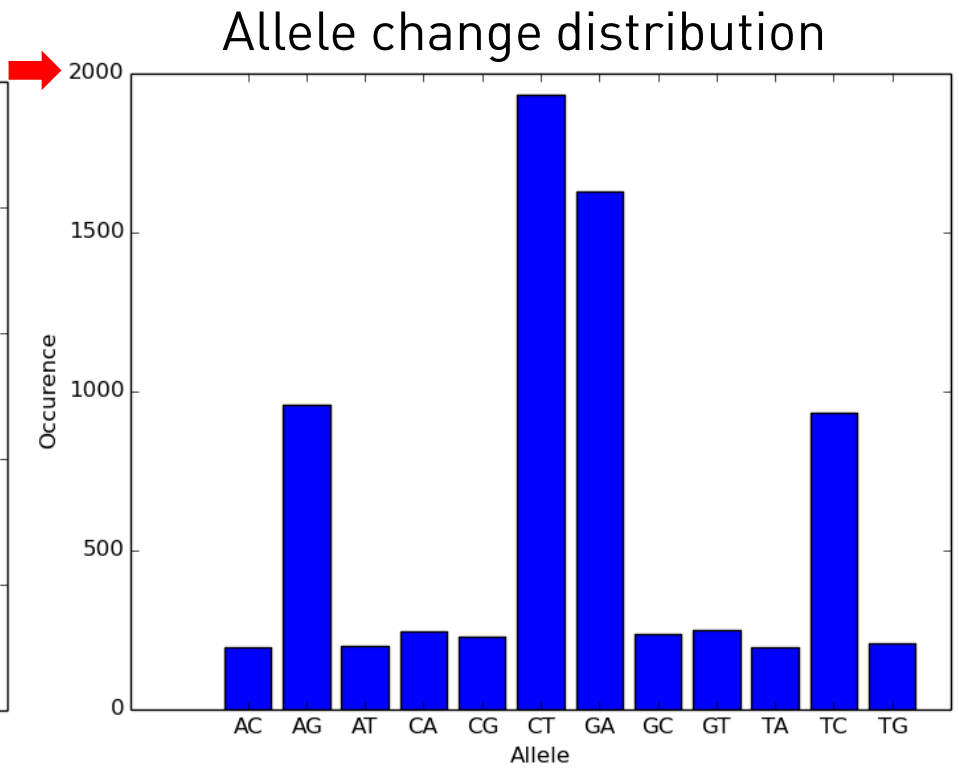


PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Damage recalibration



Raw

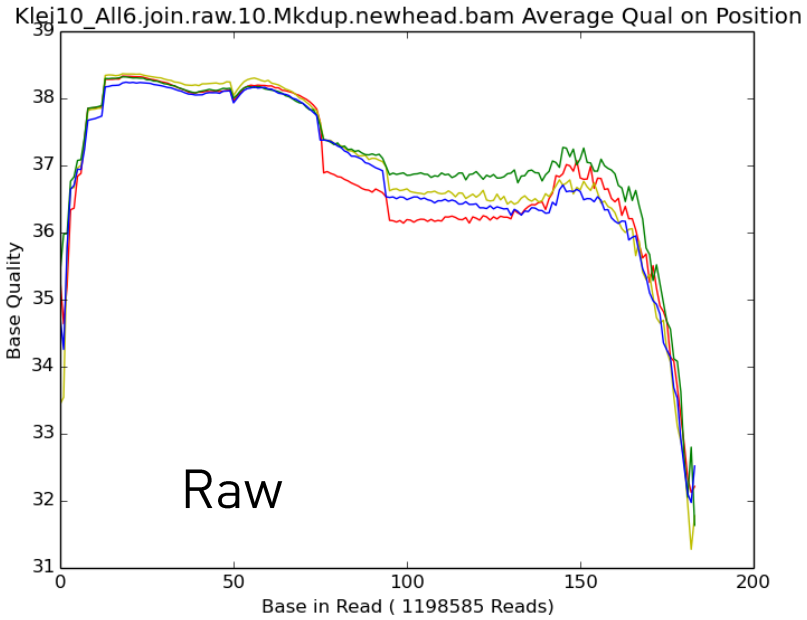


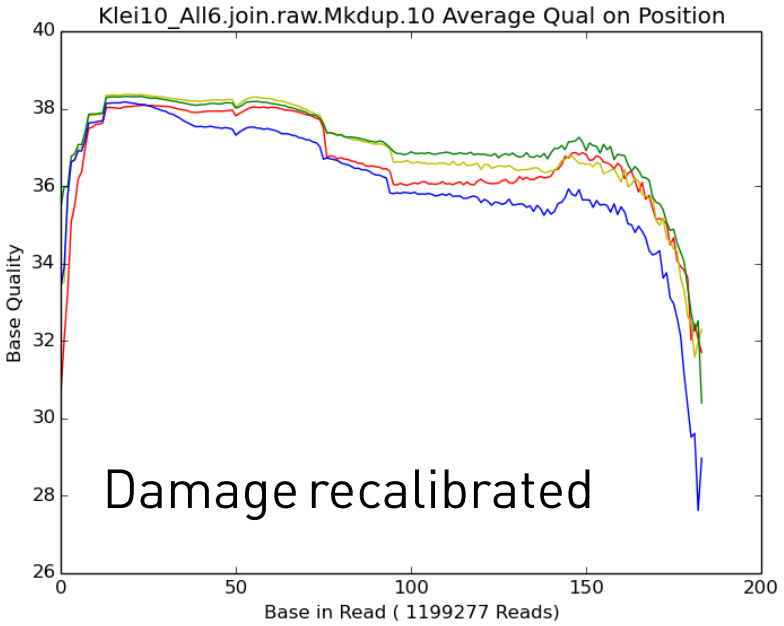
Damage recalibrated

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Damage recalibration

Average base quality/position/base



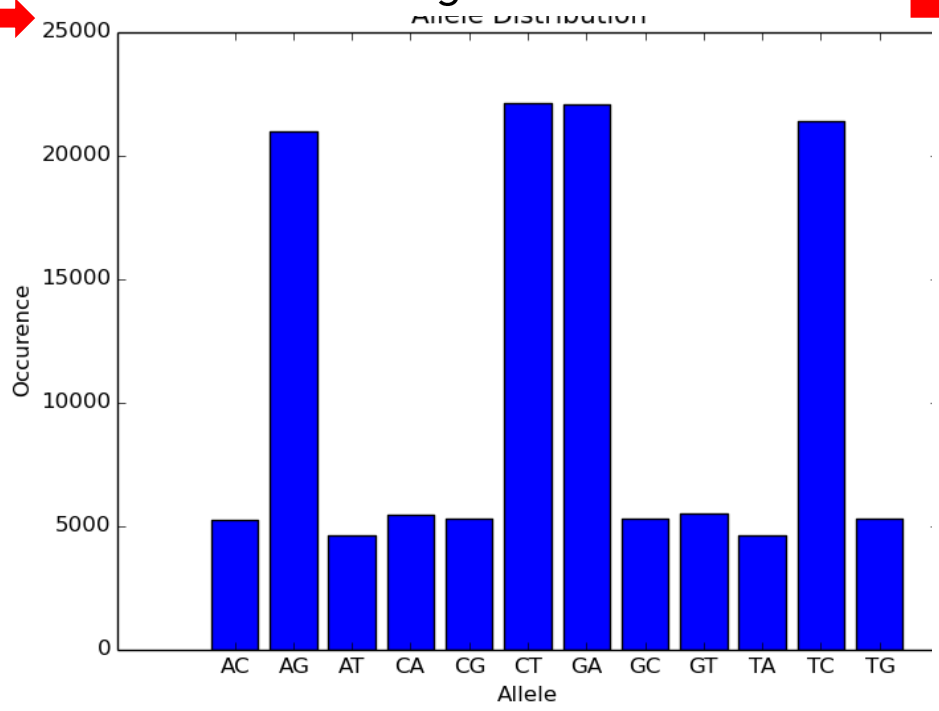


Base	Average	Min	Max
T (red)	36.86	32.13	38.33
STD	3.70	1.50	6.43
A (blue)	36.86	31.98	38.24
STD	3.48	1.69	7.31
C (yellow)	36.94	31.28	38.37
STD	3.30	1.99	7.83
G (green)	36.94	31.28	38.37
STD	3.30	1.99	7.83

Base	Average	Min	Max
T (red)	36.65	30.58	38.09
STD	4.18	2.12	11.81
A (blue)	36.13	27.62	38.17
STD	5.46	2.86	9.78
C (yellow)	36.94	31.58	38.37
STD	3.31	1.55	7.84
G (green)	36.94	31.58	38.37
STD	3.31	1.55	7.84

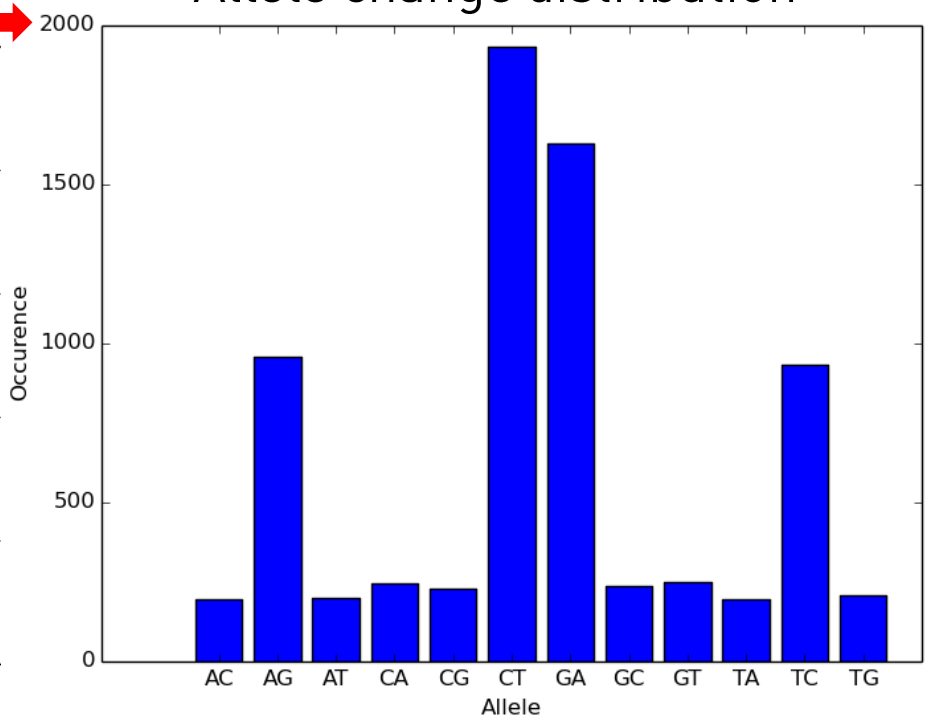
Damage recalibration

Allele change distribution



Modern

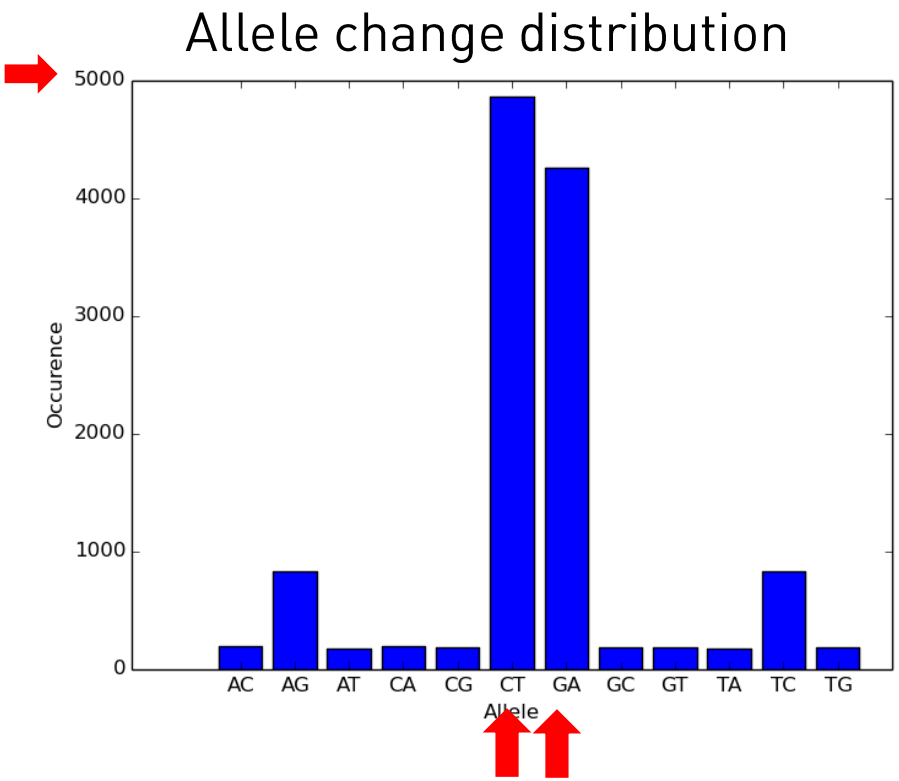
Allele change distribution



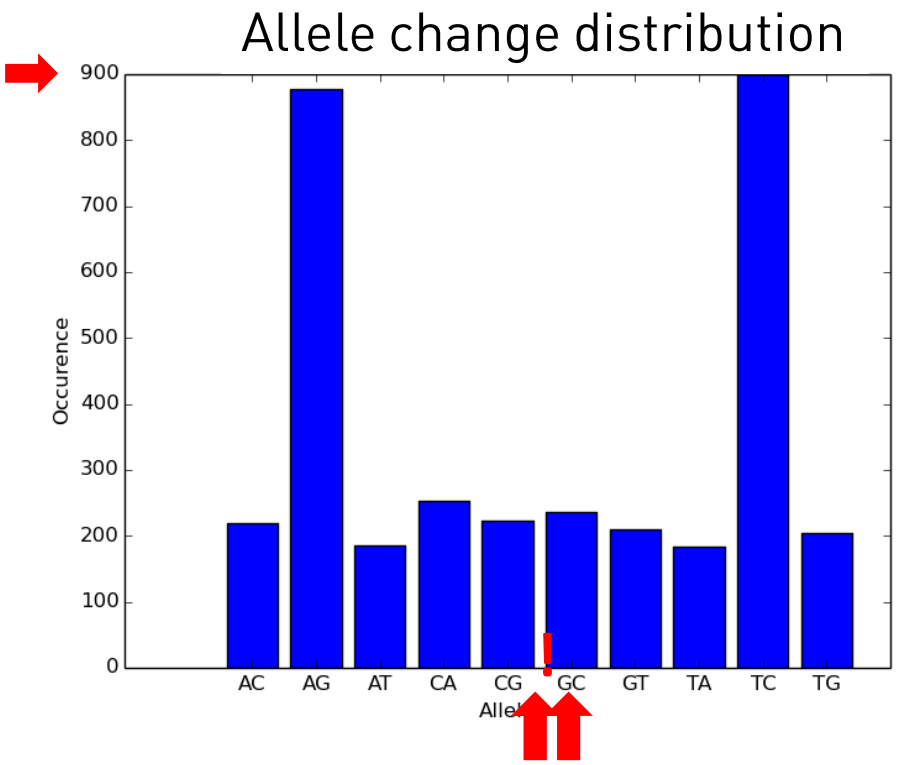
Damage recalibrated

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Damage correction followed by GATK recalibration

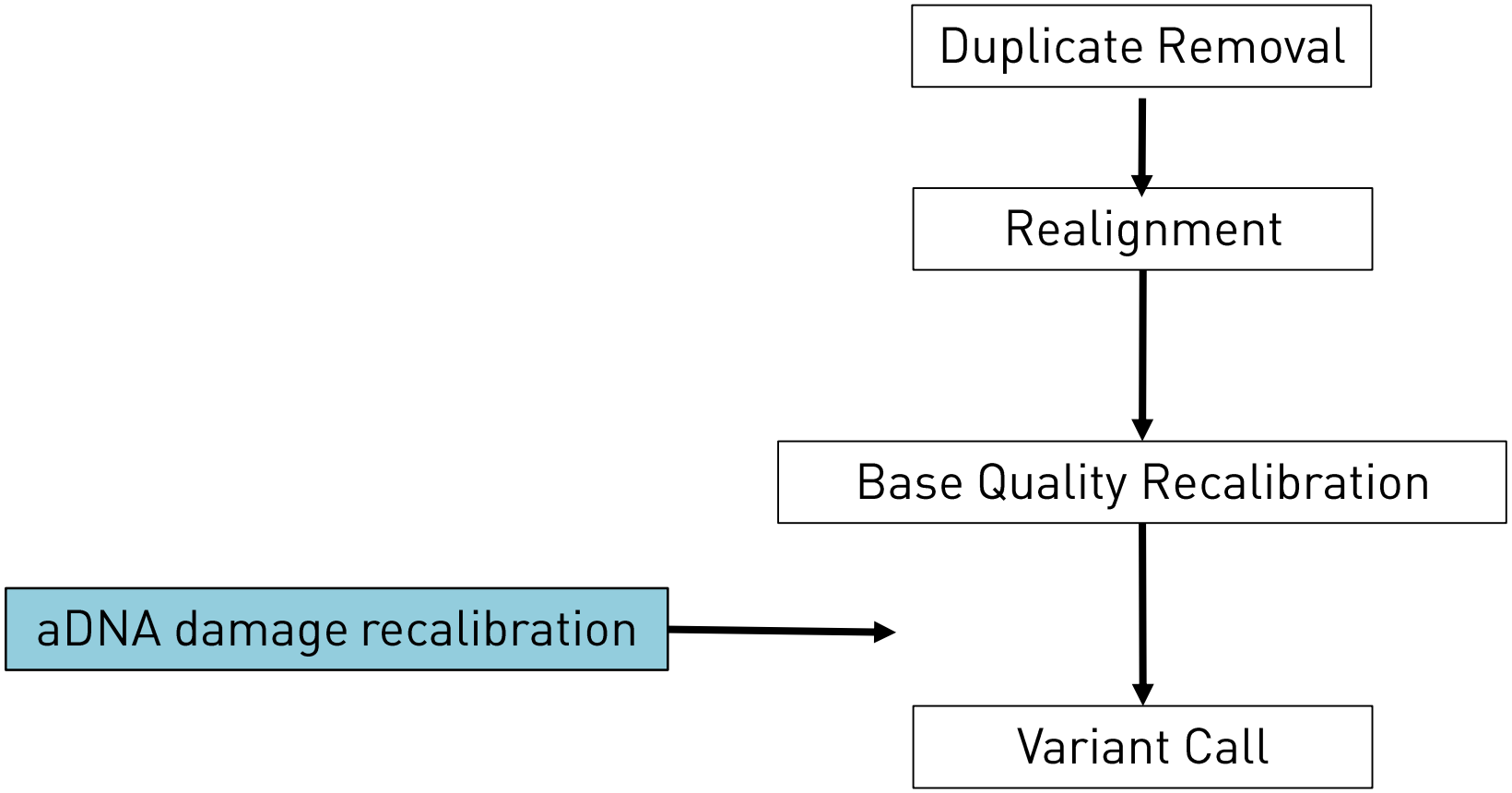


Raw



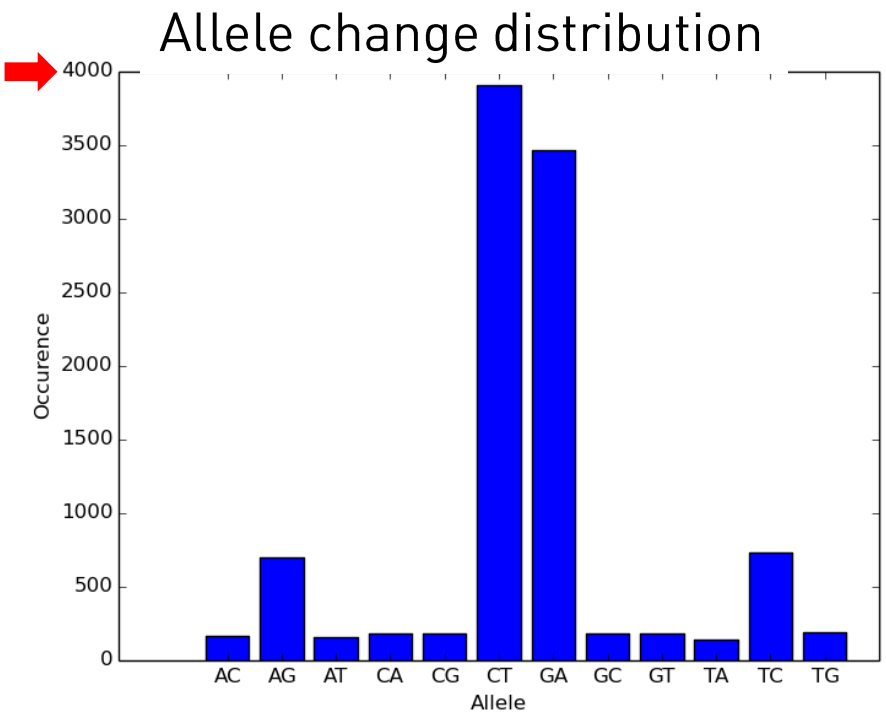
Corrected + recalibrated

Damage recalibration

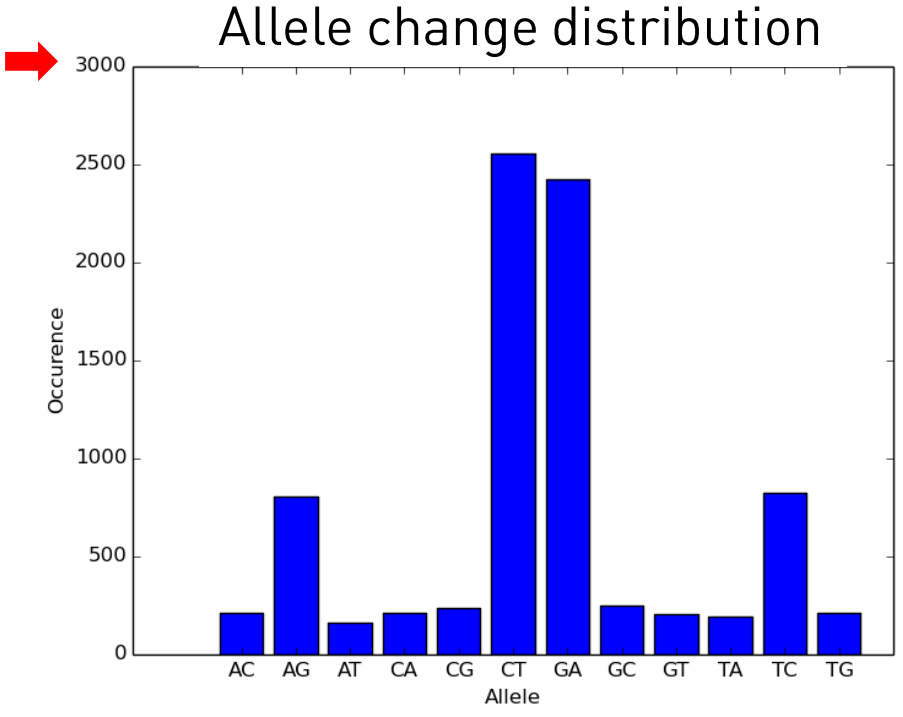


PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

GATK recalibration



Raw

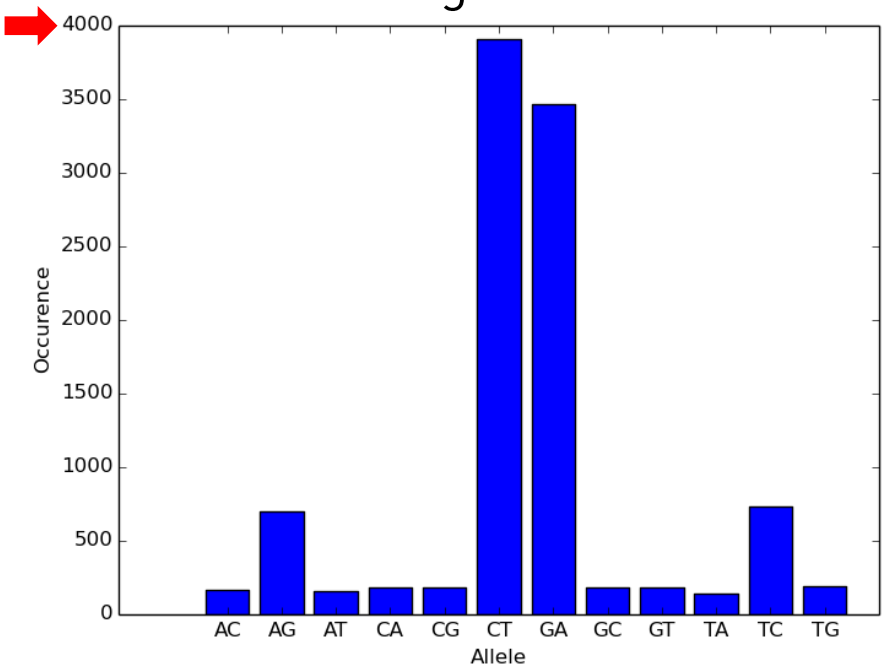


GATK recalibrated

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

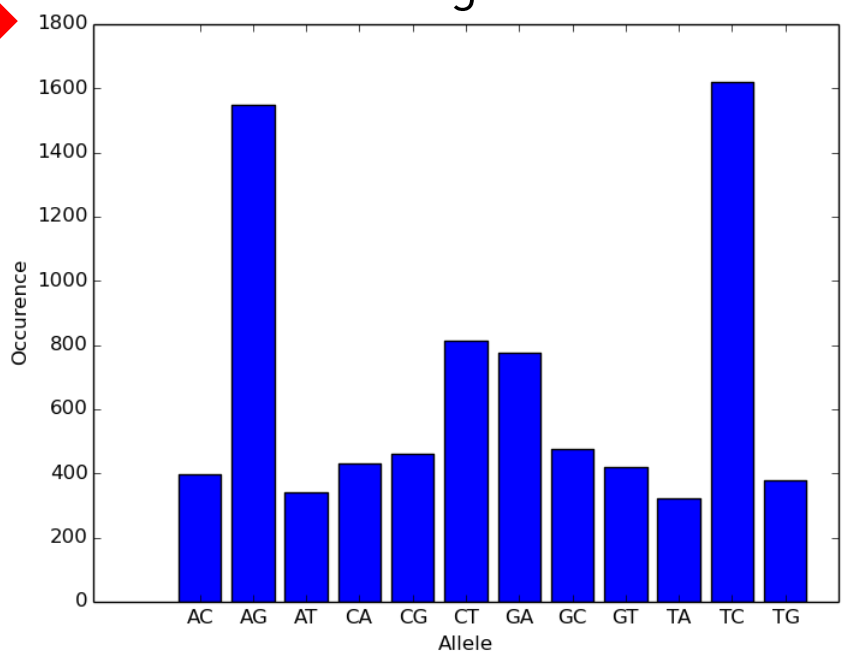
GATK recalibration followed by damage correction

Allele change distribution



Raw

Allele change distribution

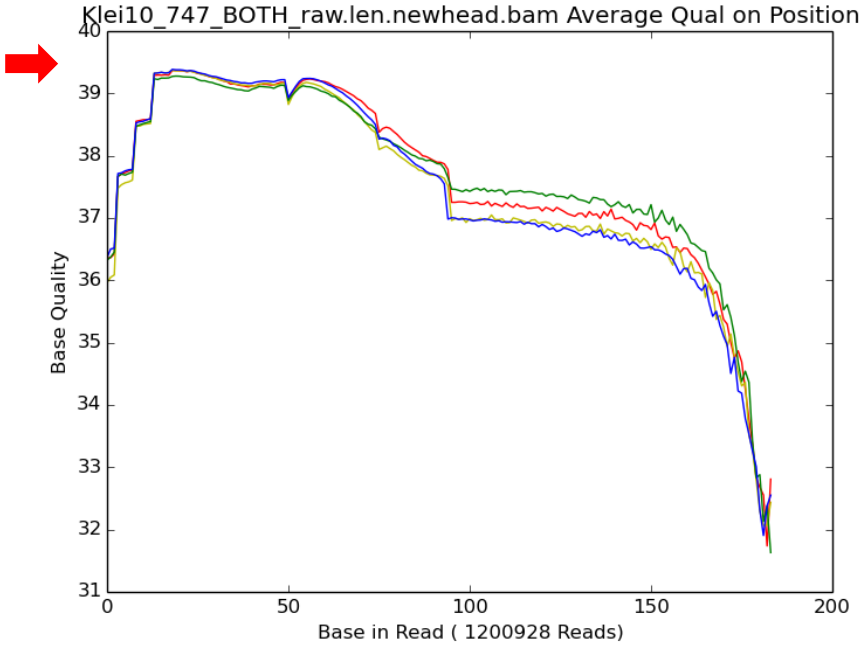


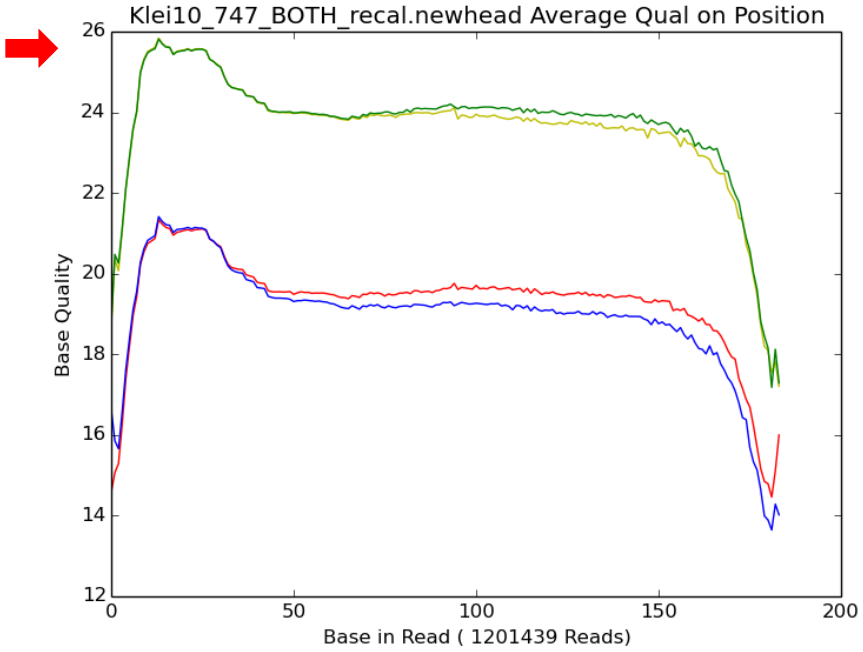
Damage and GATK recalibrated

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

GATK recal followed by damage recalibration

Average Base quality/position/base

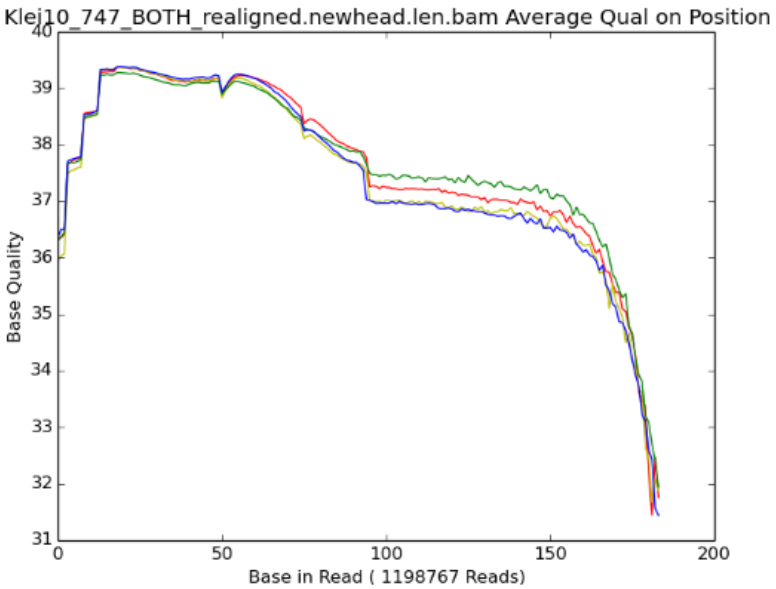


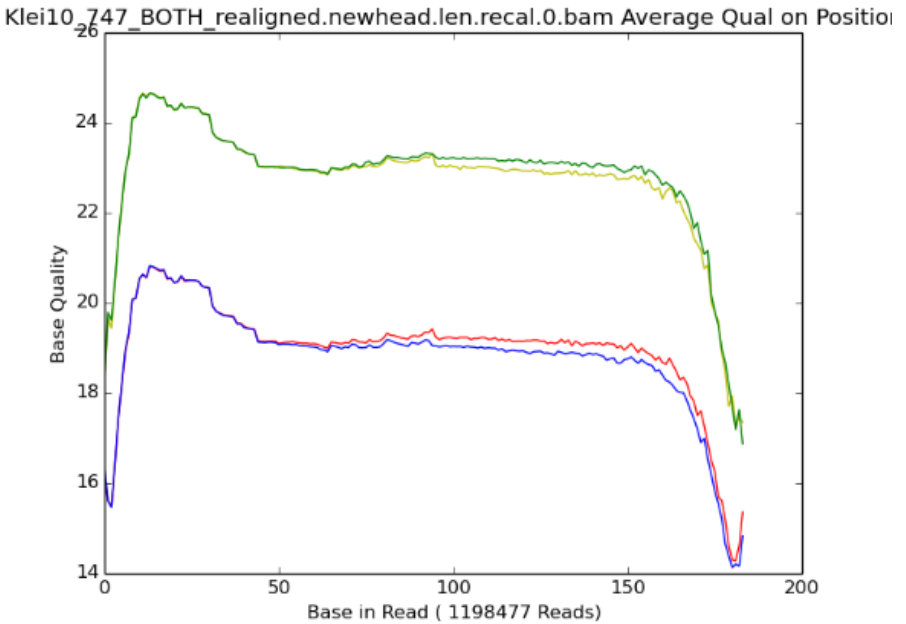


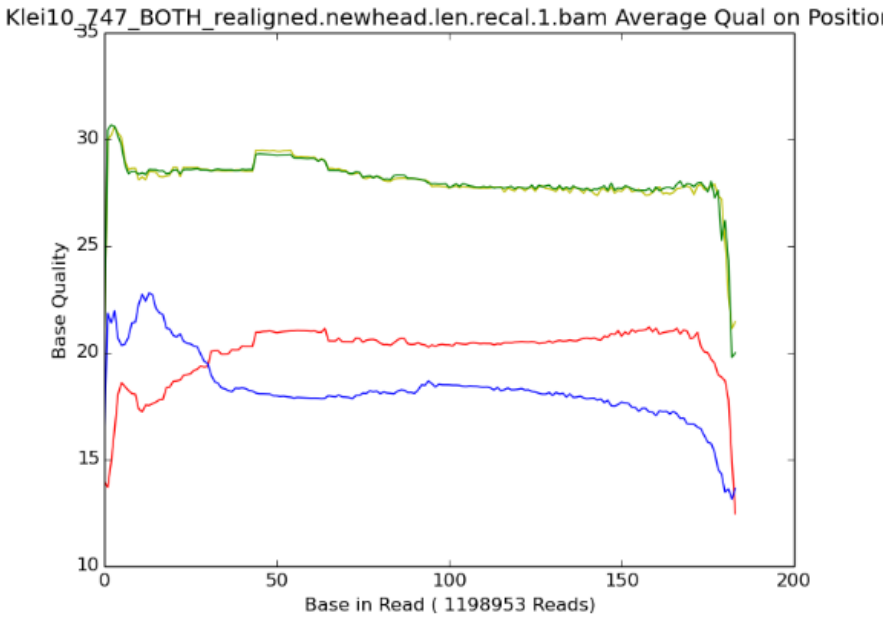
Base	Average	Min	Max
T (red)	37.63	31.74	39.37
STD	3.15	1.53	4.08
A (blue)	37.47	31.90	39.39
STD	3.23	1.58	4.21
C (yellow)	37.46	32.07	39.37
STD	3.28	1.46	4.49
G (green)	37.46	32.07	39.37
STD	3.28	1.46	4.49

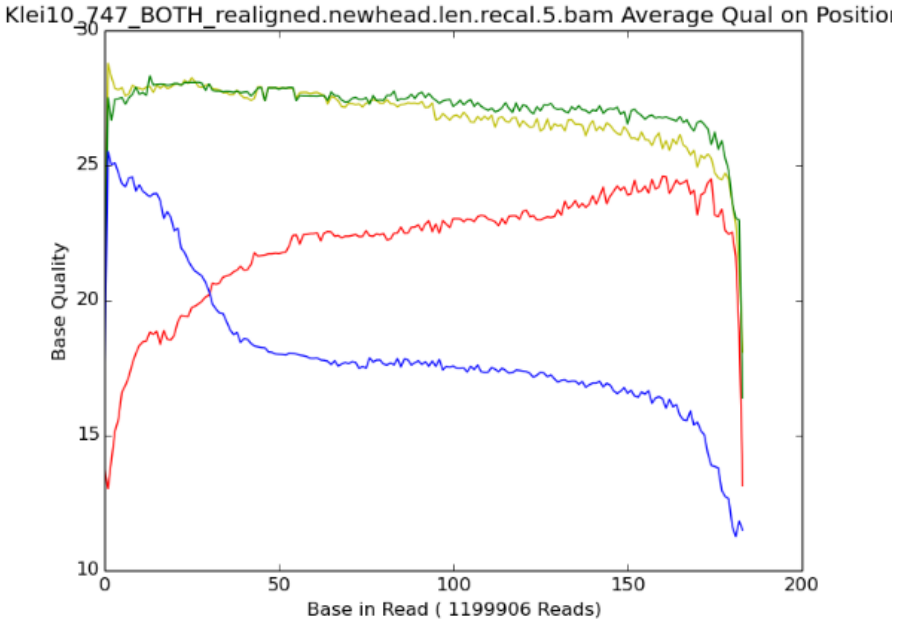
Base	Average	Min	Max
T (red)	19.29	14.46	21.34
STD	2.87	1.07	4.76
A (blue)	18.98	13.65	21.41
STD	3.27	1.99	4.14
C (yellow)	23.58	17.21	25.84
STD	2.82	1.46	4.05
G (green)	23.58	17.21	25.84
STD	2.82	1.46	4.05

GATK recalibration

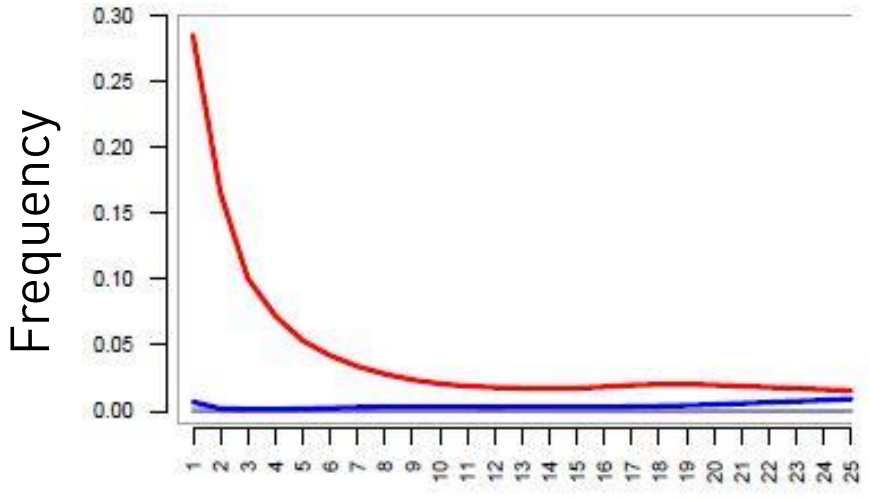




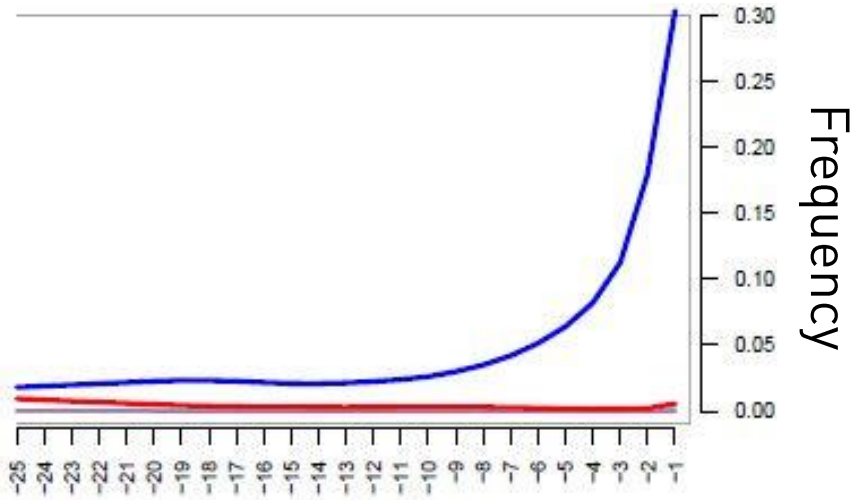




Postmortem damage patterns



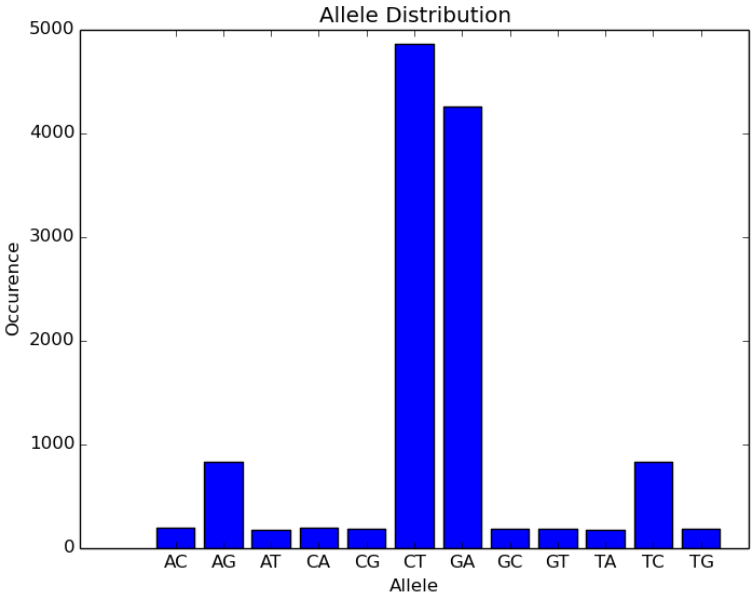
Position in read relative to 5'



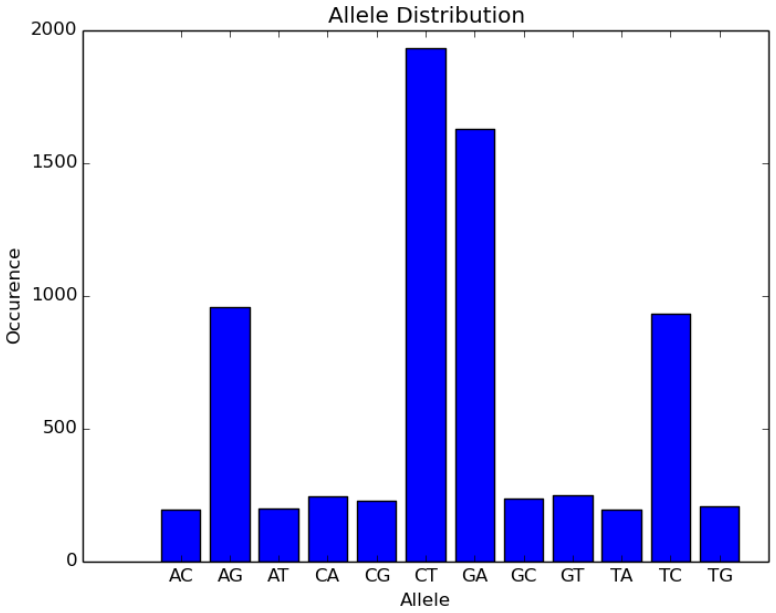
Position in read relative to 3'

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

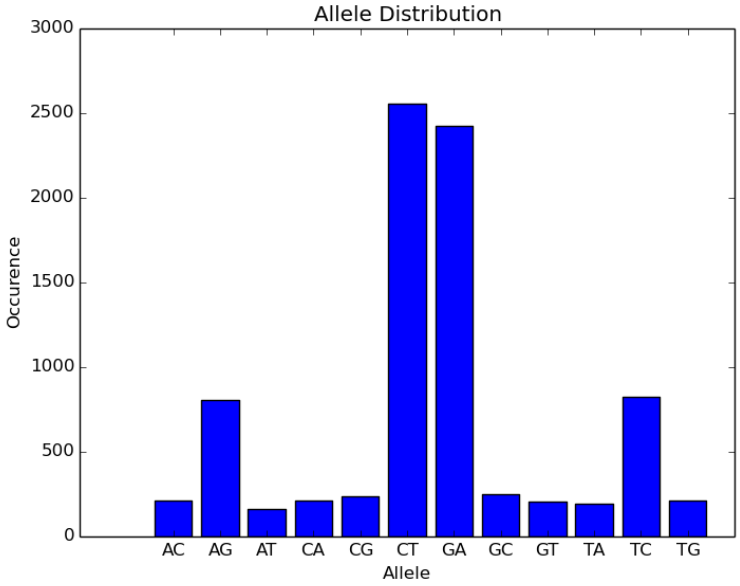
Raw data



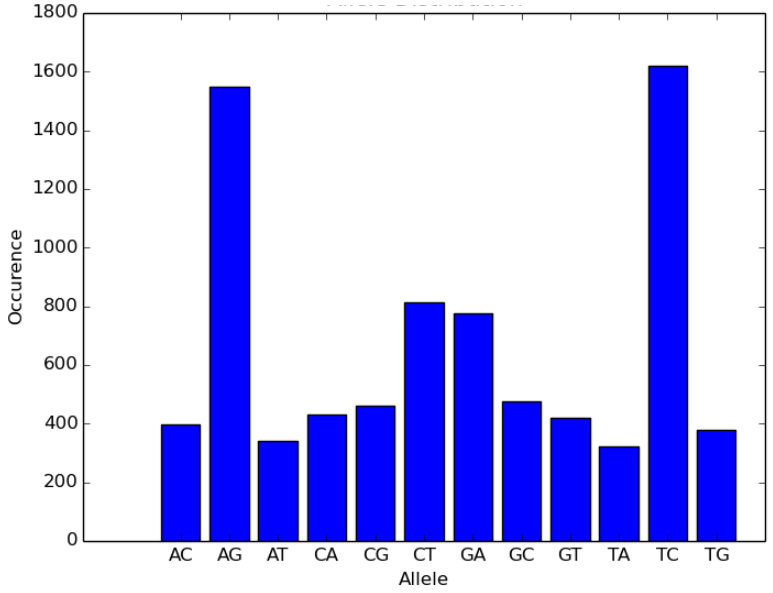
Damage recalibration



GATK recalibrated

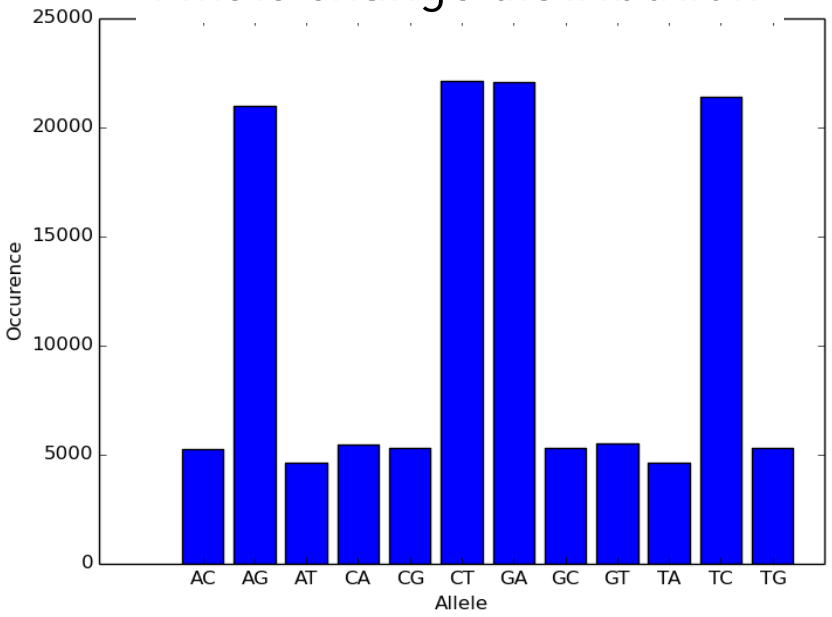


GATK recalibrated + damage recalibrated



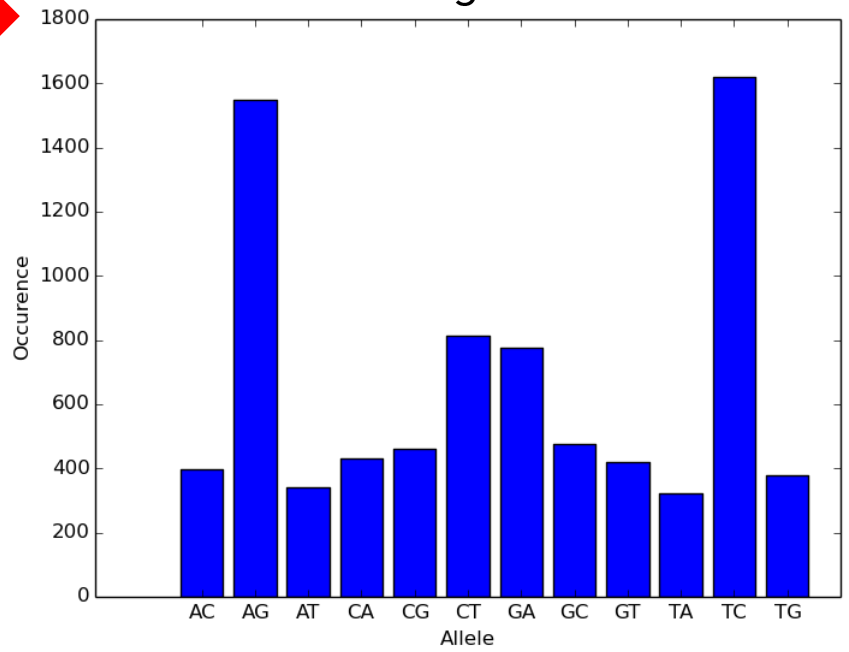
GATK recalibration followed by damage correction

Allele change distribution



Modern sample

Allele change distribution



Damage and GATK recalibrated

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

Current work

- Developing a maximum-likelihood based based SNP caller that incorporates damage correction and base recalibration into the variant calling framework
- Frees aDNA-based variant calling and genotyping from excessive dependence of modern reference
- Addresses the problem of over-training and over-correcting by running damage correction and recalibration simultaneously

PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology

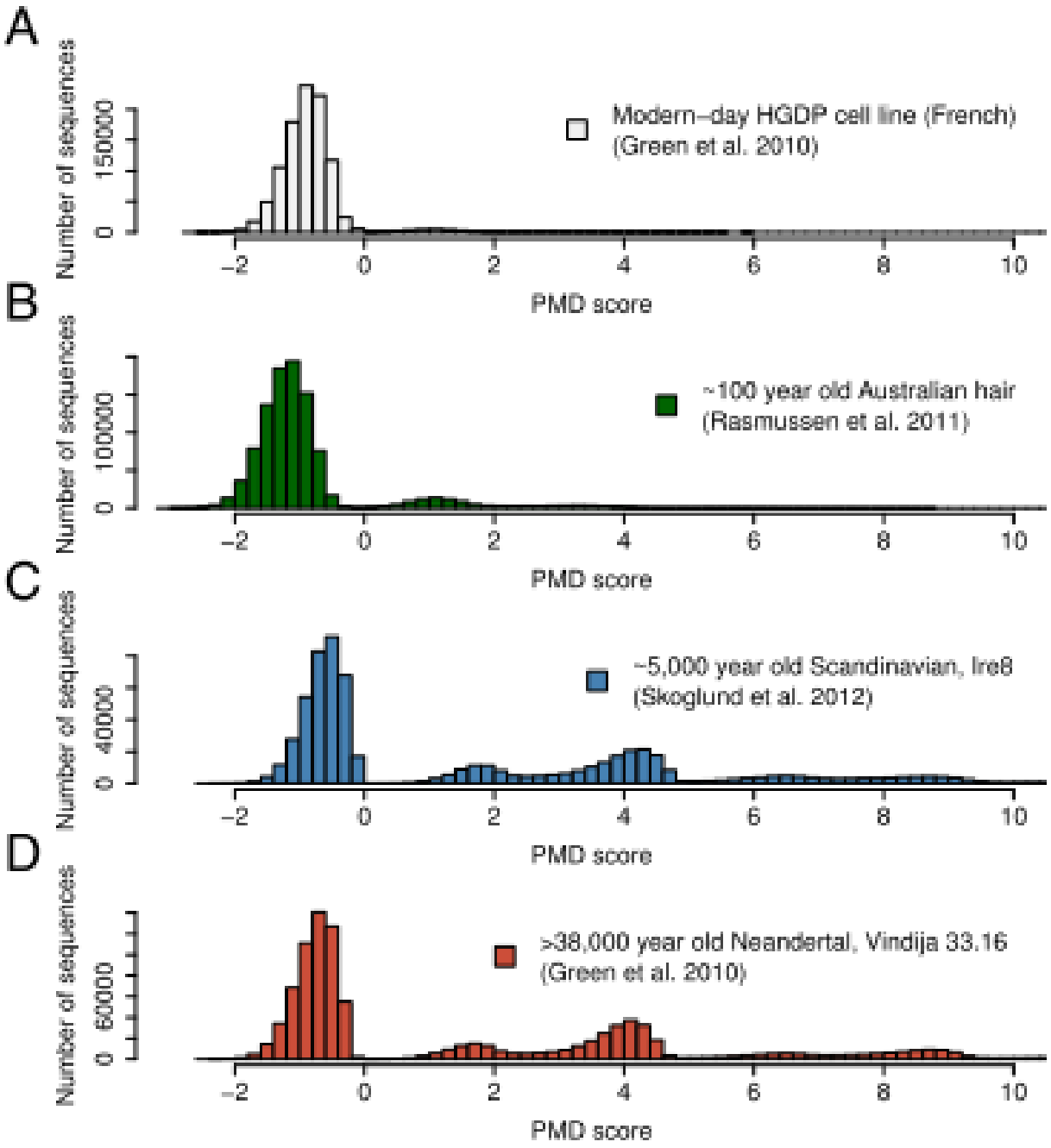
PALÆOGENETICS GROUP

at the Mainz Institute of Anthropology

STONY BROOK UNIVERSITY

GeneRED

UNIVERSITAS FRIBURGENSIS



PALÆOGENETICS GROUP
at the Mainz Institute of Anthropology