ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΧΗΜΕΙΑΣ

ΓΕΝΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΧΗΜΕΙΑΣ

ΜΟΥΣΕΙΟ ΦΥΣΙΚΗΣ ΙΣΤΟΡΙΑΣ, ΚΟΠΕΓΧΑΓΗ, ΔΑΝΙΑ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΜΕΘΟΔΩΝ ΕΜΠΛΟΥΤΙΣΜΟΥ ΓΙΑ ΤΗΝ ΑΝΑΚΤΗΣΗ ΓΕΝΕΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΑΡΧΑΙΑ ΔΕΙΓΜΑΤΑ

ΔΩΡΟΘΕΑ Ι. ΜΥΛΟΠΟΤΑΜΙΤΑΚΗ

ΥΠΕΥΘΥΝΟΣ ΚΑΘΗΓΗΤΗΣ: ΙΩΑΝΝΗΣ ΠΑΥΛΙΔΗΣ

ΗΡΑΚΛΕΙΟ 2019

UNIVERSITY OF CRETE

DEPARTMENT OF CHEMISTRY

MASTER OF CHEMISTRY

NATURAL HISTORY MUSEUM, COPENHAGEN, DENMARK

MASTER THESIS

OPTIMIZING METHODS FOR THE RECOVERY OF ANCIENT DNA SEQUENCE DATA

DOROTHEA I. MYLOPOTAMITAKI

MASTER THESIS SUPERVISOR: DR. IOANNIS PAVLIDIS

HERAKLION 2019

MASTER COMMITTEE

DR. IOANNIS PAVLIDIS

ASSISTANT PROFESSOR, UNIVERSITY OF CRETE, DEPARTMENT OF CHEMISTRY

DR. DIMITRIOS KAFETZOPOULOS

RESEARCH DIRECTOR, INSTITUTE OF MOLECULAR BIOLOGY AND BIOTECHNOLOGY, F.O.R.T.H.

DR. HANNES SCHROEDER

ASSISTANT PROFESSOR, SECTION FOR EVOLUTIONARY GENOMICS, DEPARTMENT OF BIOLOGY, UNIVERSITY OF COPENHAGEN

ACKNOWLEDGMENTS

This MSc thesis was performed at the Natural History Museum, Denmark, from September 2018 till July 2019.

To begin with, I would like to thank Dr. Hannes Schroeder, Section for Evolutionary Genomics, University of Copenhagen, Denmark, for giving me the opportunity to accomplish my research as part of his group and for supporting my interest on this field.

Additionally, I would like to thank Dr. Dimitrios Kafetzopoulos, Institute of Molecular Biology and Biotechnology (IMBB) Foundation of Research (FORTH) Crete, and Dr. Ioannis Pavlidis, University of Crete, Department of Chemistry, for being members of my master's committee and agreeing on evaluating this report.

I am also grateful to all the people that helped on this project. Specifically, Alba Rey de la Iglesia (National History Museum, Denmark) for sharing her knowledge and experience on this project with me, Shyam Gopalakrishnan and Claudia Sarai Reyes Avila (University of Copenhagen, Denmark) for discussions and guidance on data analysis, Tobias Mourier (King Abdullah University of Science and Technology, Saudi Arabia) and Shyam Gopalakrishnan for designing the Alu primers, Kristian Murphy Gregersen (Natural History Museum, Denmark) who provided samples for this study and Anders Johannes Hansen (Natural History Museum, Denmark) for his help and guidance.

I would like to thank the Erasmus+ mobility programme for EU students for partly financial support of my mobility; Mrs. Eugenia Skountaki as the Erasmus+ students coordinator, School of Science & Engineering, University of Crete, Heraklion Campus and Prof. Haralambos E. Katerinopoulos, University of Crete, Department of Chemistry for their help.

Finally, I cannot thank enough all the members in our research group for their advice and support and my family and friends for their support over that period.

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ

ΠΡΟΣΩΠΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ

Ονοματεπώνυμο: **Δωροθέα Μυλοποταμιτάκη**
Ημερομηνία γέννησης: 7 Φεβρουαρίου 1994
Τόπος γέννησης: Ηράκλειο, Κρήτη, Ελλάδα.
Φύλο: Θηλυκό

Διεύθυνση: Αλεξάνδρου Παπαναστασίου 230, Ηράκλειο, Κρήτη, Ελλάδα, 71409
E-mail: dorothymilop@gmail.com
Τηλέφωνο: **0030 6970808459**

ΕΚΠΑΙΔΕΥΣΗ

- **2017- σήμερα:** Μεταπτυχιακή φοιτήτρια στο Τμήμα Χημείας του Πανεπιστημίου Κρήτης, Ελλάδα
- **2012- 2017:** Φοιτήτρια στο Τμήμα Χημείας του Πανεπιστημίου Κρήτης, Ελλάδα (Βαθμός αποφοίτησης: 6.85/10)
- **2007- 2010:** Αποφοίτηση από το 3ο Γενικό Λύκειο Ηρακλείου Κρήτης, Ελλάδα (Βαθμός αποφοίτησης: 18.5/20)

Τίτλοι Ερευνητικών Εργασιών:

- **Μεταπτυχιακή Διατριβή**
  Σεπτέμβριος 2018- Ιούλιος 2019: *"Βελτιστοποίηση μεθόδων ανάκτησης γενετικών δεδομένων από αρχαία δείγματα"* υπό την επίβλεψη του Δρ. Hannes Schroeder, Πανεπιστήμιο της Κοπεγχάγης, Τμήμα Βιολογίας, Δανία.
- **Πτυχιακή Ερευνητική Εργασία**

Νοέμβριος 2016- Ιούνιος 2017: *«Βασικές Μελέτες για την Επεξεργασία των Αρχαιολογικών Δειγμάτων και την Ανάκτηση Αρχαίου DNA για περαιτέρω Ανάλυση»* υπό την επίβλεψη του Διευθυντή Ερευνών Δ. Καφετζόπουλο, Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας (IMBB), Ίδρυμα Τεχνολογίας και Έρευνας (ITE) Κρήτη.

ΕΚΠΑΙΔΕΥΤΙΚΗ ΕΜΠΕΙΡΙΑ

- **Νοέμβριος 2017- Δεκέμβριος 2017:** Διδάσκων στο "Ancient DNA & Evolution, Laboratory Practicals" σε προπτυχιακούς φοιτητές του τμήματος Βιολογίας στο Πανεπιστήμιο της Κοπεγχάγης. Το εκπαιδευτικό πρόγραμμα πραγματοποιήθηκε στα εργαστήρια του Μουσείο Φυσικής Ιστορίας της Δανίας με σκοπό οι φοιτητές να αποκτήσουν εργαστηριακή εμπειρία σε πρωτόκολλα απομόνωσης και τεχνικές σχετικές με την ανάλυση αρχαίου DNA.

ΕΡΕΥΝΗΤΙΚΗ ΕΜΠΕΙΡΙΑ

- **Σεπτέμβριος 2018- παρόν:** Πτυχιούχος ασκούμενος, 10-μηνη πρακτική άσκηση μέσω υποτροφίας για πρακτική *Erasmus*$^{+}$. Τίτλος άσκησης *"Βελτιστοποίηση μεθόδων ανάκτησης γενετικών δεδομένων από αρχαία δείγματα"* στο Center of Geogenetics υπό την επίβλεψη του Δρ. Hannes Schroeder, Πανεπιστήμιο της Κοπεγχάγης, Τμήμα Βιολογίας, Δανία.
- **Ιούλιος 2017- Σεπτέμβριος 2017:** Τρίμηνη πρακτική άσκηση μέσω υποτροφίας για πρακτική *Erasmus*$^{+}$. Τίτλος άσκησης "*High-throughput sequencing applications in archaeology"* στο Center of Geogenetics υπό την επίβλεψη του Δρ. Hannes Schroeder, Πανεπιστήμιο της Κοπεγχάγης, Τμήμα Βιολογίας, Δανία.

- **Νοέμβριος 2016- Ιούνιος 2017:** Προπτυχιακή ασκούμενη στο Εργαστήριο Ανάλυσης Αρχαίου DNA του IMBB-ITE με Επιστημονικό Υπεύθυνο τον Διευθυντή Ερευνών Δ. Καφετζόπουλο.

ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΕΜΠΕΙΡΙΑ

- **Οκτώβριος 2017- Μάρτιος 2018:** Τεχνικός επιμελητής στη ερευνητική πρόγραμμα "*The genomic history of Trondheim" στο Centre of Geogenetics* υπό την επίβλεψη του καθηγητή M. T. P. Gilbert στον τομέα Evolutionary Genomics, Πανεπιστήμιο της Κοπεγχάγης, Δανία.

ΜΑΘΗΜΑΤΑ ΚΑΤΑΡΤΙΣΗΣ

- **Ιούλιος 2018:** Παρακολούθηση ως επισκέπτης το **"**Archaeomics summer school" στο Πανεπιστήμιο της Κοπεγχάγης, Δανία.

ΔΕΞΙΟΤΗΤΕΣ ΚΑΙ ΕΠΑΡΚΕΙΑ

- <u>Αναλυτικές τεχνικές</u>
  - Χρωματογραφία λεπτού στιβάδος (TLC)
  - UV/VIS Φασματοσκοπία
  - Χρωματογραφία στήλης

- <u>Τεχνικές μικροσκοπίας</u>
  - Ηλεκτρονική μικροσκοπία

- <u>Μοριακές τεχνικές</u>
  - Πρωτόκολλα απομόνωσης DNA

- DNA βιβλιοθήκες
- Πρωτόκολλα στοχευμένου πολλαπλασιασμού DNA
- PCR, Real- time PCR
- Φθοριομετρική ποσοτικοποίηση DNA (Bioanalyzer and TapeStation)
- Ηλεκτροφόρηση σε πηκτή

▪ Ξένες γλώσσες
- Αγγλικά: IELTS
- Γερμανικά: Goethe-Zertifikat B2

▪ Γνώση υπολογιστών
- Microsoft Office
- OriginPro
- Adobe Photoshop

ΣΥΝΕΔΡΙΑ

- **Μάρτιος 2016:** 18th Συνέδριο Μεταπτυχιακών φοιτητών Χημείας, Τμήμα Χημείας, Πανεπιστημίου Κρήτης, Ηράκλειο, Κρήτη.

- **Μάιος 2016:** Παγκόσμιο Επιστημονικό Συνέδριο με τίτλο “Ανάλυση αρχαίου DNA: Μια νέα ματιά στο παρελθόν”, Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας (IMBB), Ίδρυμα Τεχνολογίας και Έρευνας (ITE), Ηράκλειο, Κρήτη.

- **Μάρτιος 2017:** COST Annual Meeting 2017: Native Mass spectrometry and related methods in Structural Biology, Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας (IMBB), Ίδρυμα Τεχνολογίας και Έρευνας (ITE), Ηράκλειο Κρήτη.

- **Μάιος 2017:** 19ο Συνέδριο Μεταπτυχιακών φοιτητών Χημείας, Τμήμα Χημείας, Πανεπιστημίου Κρήτης, Ηράκλειο, Κρήτη.

ΔΗΜΟΣΙΕΥΣΕΙΣ

Επιστημονικά άρθρα:

- Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, Rasmussen S, Margaryan A, Iraeta Orbegozo M, **Mylopotamitaki D**, Wooller M, Bataille C, Becerra-Valdivia L, Chivall D, Comeskey D, Devièse T, Grayson DK, George L, Harry H, Alexandersen V, Primeau C, Erlandson J, Rodrigues-Carvalho C, Reis S, Bastos MQR, Cybulski J, Vullo C, Morello F, Vilar M, Wells S, Gregersen K, Hansen KL, Lynnerup N,Mirazón Lahr M, Kjær K, Strauss A, Alfonso-Durruty M, Salas A, Schroeder H, Higham T, Malhi RS, Rasic JT, Souza L, Santos FR, Malaspinas AS, Sikora M, Nielsen R, Song YS, Meltzer DJ, Willerslev E., (2018), Early Human Dispersals within the Americas. *Science, 362, (6419).* http://dx.doi.org/10.1126/science.aav2621

CURRICULUM VITAE

PERSONAL DETAILS

Full name: **Dorothea MYLOPOTAMITAKI**
Date of birth: February 7th, 1994
Place of birth: Heraklion, Greece.
Gender: Female

Home address: Alexander Papanastasiou 230, 71409, Heraklion, Greece.
E-mail: dorothymilop@gmail.com
tel.: **0030 6970808459**

EDUCATION

- **2017- present:** Master student in the Department of Chemistry, University of Crete, Greece.
- **2012- 2017:** Undergraduate student in the Department of Chemistry, University of Crete, Greece (grade: 6.85/10)
- **2007- 2010:** 3rd General High School, Heraklion, Crete. Graduated with Honors (grade: 18.5/20)

Thesis Titles:

- **Master Thesis**
  September 2018- July 2019: *"Optimizing methods for the recovery of ancient DNA sequence data"* under the supervision of Dr. Hannes Schroeder, Section for Evolutionary Genomics, University of Copenhagen, Denmark, Denmark.

- **Bachelor Thesis**

  November 2016- June 2017: "*Basic studies for processing archaeological samples and the recovery of ancient DNA for further analysis*", under the supervision of Prof. D. Kafetzopoulos, Laboratory of Ancient DNA, Institute of Molecular Biology and Biotechnology (IMBB) Foundation of Research (FORTH) Crete.

EDUCATIONAL EXPERIENCE

- **November 2017- December 2017:** Tutor of "Ancient DNA & Evolution, Laboratory Practicals" to bachelor students of the Biology Department, University of Copenhagen. The tutoring took place in the Natural History Museum of Denmark and performed in order the students to obtain lab skills in the extraction protocols and techniques of ancient DNA.

RESEARCH EXPERIENCE

- **September 2018- present:** Graduate trainee, 10-months traineeship as an *Erasmus*$^{+}$ trainee. Placement title *"Optimizing methods for the recovery of ancient DNA sequence data"* in the Center of Geogenetics under the supervision of Dr. Hannes Schroeder, Section for Evolutionary Genomics, University of Copenhagen, Denmark.
- **July 2017- September 2017:** 3-months traineeship as an *Erasmus*$^{+}$ trainee. Placement title "*High-throughput sequencing applications in archaeology"* in the Center of Geogenetics under the supervision of Dr. Hannes Schroeder, Section for Evolutionary Genomics, University of Copenhagen, Denmark.
- **November 2016- June 2017:** Undergraduate trainee Undergraduate trainee in the Laboratory of Ancient DNA Institute of Molecular Biology

and Biotechnology (IMBB), Foundation of Research and Technology (FORTH) Crete under the supervision PI Dr. D. Kafetzopoulos

WORKING EXPERIENCE

- **October 2017- March 2018:** Technical assistant on the project "*The genomic history of Trondheim" in the Centre of Geogenetics* under the supervision of Professor M. T. P. Gilbert in the Section for Evolutionary Genomics, University of Copenhagen, Denmark.

TRAINING COURSES

- **July 2018:** Attended as a visitor in Archaeomics summer school at the University of Copenhagen, Denmark.

SKILLS AND COMPETENCE

- Analytical Techniques
  - Thin layer chromatography (TLC)
  - UV/VIS Spectrophotometer
  - Column chromatography

- Microscopy Techniques
  - Electron Microscopy

- Molecular techniques
  - DNA extraction protocols
  - Library buildings
  - Targeting DNA protocols
  - PCR, Real- time PCR
  - Fluorometric quantitation of DNA (Bioanalyzer and TapeStation)

- Gel electrophoresis

- Foreign languages
  - English: IELTS
  - German: Goethe-Zertifikat B2

- Computer skills
  - Microsoft Office
  - OriginPro
  - Adobe Photoshop

CONFERENCES

- **March 2016:** 18th Postgraduates' conference on Chemistry, Department of Chemistry University of Crete, Heraklion, Crete, Greece.
- **May 2016:** International Scientific Conference titled "Analysis of Ancient DNA: A New Matter in the Past", Institute of Molecular Biology and Biotechnology (IMBB) of the Foundation for Research and Technology (FORTH), Heraklion, Crete, Greece.
- **March 2017:** COST Annual Meeting 2017: Native Mass spectrometry and related methods in Structural Biology, Institute of Molecular Biology and Biotechnology (IMBB) of the Foundation for Research and Technology (FORTH), Heraklion, Crete, Greece.
- **May 2017:** 19th Postgraduates' conference on Chemistry, Department of Chemistry, University of Crete, Heraklion, Greece.

PUBLICATIONS

Journal articles:

Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, Rasmussen S, Margaryan A, Iraeta Orbegozo M, **Mylopotamitaki D**, Wooller M, Bataille C, Becerra-Valdivia L, Chivall D, Comeskey D, Devièse T, Grayson DK, George L, Harry H, Alexandersen V, Primeau C, Erlandson J, Rodrigues-Carvalho C, Reis S, Bastos MQR, Cybulski J, Vullo C, Morello F, Vilar M, Wells S, Gregersen K, Hansen KL, Lynnerup N,Mirazón Lahr M, Kjær K, Strauss A, Alfonso-Durruty M, Salas A, Schroeder H, Higham T, Malhi RS, Rasic JT, Souza L, Santos FR, Malaspinas AS, Sikora M, Nielsen R, Song YS, Meltzer DJ, Willerslev E., (2018), Early Human Dispersals within the Americas. *Science, 362, (6419).* http://dx.doi.org/10.1126/science.aav2621

ΠΕΡΙΛΗΨΗ

Δυσκολίες στην ανάκτηση δεδομένων αλληλούχισης γενετικού υλικού από αρχαιολογικά δείγματα, όπως η χαμηλή ποιότητας των δειγμάτων και η περιορισμένη ποσότητα του προ ανάλυση υλικού, έχουν μερικώς αντιμετωπιστεί με την ανάπτυξη των τεχνολογιών αλληλούχισης υψηλής απόδοσης (high-throughput sequencing) και τις μεθόδους εμπλουτισμού στοχευμένης σύλληψης (target enrichment methods). Παρόλα αυτά, οι περισσότερες τεχνικές εμπλουτισμού στοχευμένης σύλληψης απαιτούν ακριβούς ανιχνευτές (Hardenbol et al. 2005; Lizardi et al. 1998) ή έχουν περιορισμένη διαθεσιμότητα στο κοινό (Mathieson et al. 2015).

Η συγκεκριμένη μελέτη επικεντρώνεται στην ανάπτυξη μιας μεθόδου εμπλουτισμού στοχευμένης σύλληψης για αρχαίο ανθρώπινο γενετικό υλικό βασισμένη στη μέθοδο αλληλούχισης μετακινούμενων στοιχείων (Mobile elements Sequencing, MobiSeq) η οποία σχεδιάστηκε για μη ανθρώπινα χαμηλής ποιότητας δείγματα (Rey-Iglesia, Gopalakrishnan, and Carøe 2018). Η μέθοδος εμπλουτισμού στοχευμένης σύλληψης MobiSeq βασίζεται στον εμπλουτισμό του στόχου μέσω PCR αντίδρασης και χρησιμοποιεί απλές εργαστηριακές τεχνικές για την ανάκτηση δεδομένων αλληλούχισης από αρχαιολογικά δείγματα. Η συγκεκριμένη μέθοδος χρησιμοποιεί μεταθετά στοιχεία (transposable elements, TEs) ως “άγκυρα” για την αλληλούχιση των πλευρικών περιοχών των TEs.

Αξιολογήσαμε την επίδοση της συγκεκριμένης μεθόδου ανάμεσα σε 20 εκχυλίσματα DNA από ανθρώπινα δείγματα με τη βοήθεια δύο διαφορετικών TE εκκινητών για οικογένειες Alu γονιδίων, οι οποίες δίνουν σημάδια πρόσφατης δραστηριότητας. Χρησιμοποιήθηκαν, επίσης, βιβλιοθήκες τύπου “shotgun” με σκοπό τη σύγκριση του αριθμού των μονο-νουκλεοτιδικών πολυμορφισμών (SNPs) που μπορούν να ανακτηθούν από τις δύο παραπάνω μεθόδους. Η τεχνική εμπλουτισμού στοχευμένης σύλληψης MobiSeq σε μη ανθρώπινα και χαμηλής ποιότητας δείγματα έδωσε 90% ανάκτηση SNPs με χαμηλές τιμές κλωνικότητας. Παρόλο που ο Alu_v1 TE εκκινητής έδειξε αυξημένη ανάνηψη SNP για 4 ανθρώπινα δείγματα, βελτιστοποίηση της μεθόδου για την ανακάλυψη περισσότερων SNPs μπορεί να πραγματοποιηθεί με αλλαγή του TE εκκινητή, καλύτερη ποσοτική ανάλυση των δειγμάτων, αλλαγή της

αλληλούχισης και της βιοπληροφορικής ανάλυσης. Περαιτέρω βιοπληροφορική μελέτη απαιτείται για την πλήρη αξιολόγηση της συγκεκριμένης τεχνικής.

Λέξεις κλειδιά: αρχαίο DNA, next-generation sequencing, εμπλουτισμός στοχευμένης γονιδιακής αλληλουχίας, μεταθετά γονιδιωματικά στοιχεία, γονιδιωματική πληθυσμών, ανακάλυψη SNPs.

ABSTRACT

Challenges in ancient DNA (aDNA) research such as limited quantity of starting material and highly degraded DNA are partially encountered using high-throughput sequencing approaches and targeted enrichment (capture) methods for recovering sequence data. Nevertheless, most of the techniques focused on target enrichment require expensive probes (Hardenbol et al. 2005; Lizardi et al. 1998) or have limited availability to the public (Mathieson et al. 2015).

This study aims to develop a target enrichment method for ancient human DNA based on a Mobile element Sequencing (MobiSeq) Reduced Representation Library (RRL) protocol which was originally designed for non-human degraded samples (Rey-Iglesia, Gopalakrishnan, and Carøe 2018). MobiSeq approach is a PCR-based targeted-enrichment method generating sequence data from ancient human samples using simple laboratory techniques. This method uses transposable elements (TEs) as “anchor” for sequencing extension into the flanking region of these mobile elements.

To evaluate the performance of this method 20 human DNA extracts were targeted by two different TE-target primers for Alu gene families that showed hallmarks of recent activity. Shotgun libraries were also built to compare the number of single-nucleotide polymorphisms (SNPs) that could be recovered from 1k genomes with our method. MobiSeq target enrichment method generated 90% of loci across the genome and performs SNP discovery with relatively low rates of clonality in non-human species. Although, Alu_v1 TE-target primer showed high SNP discovery for 4 of the human samples, optimization of this method can be performed on TE-target primer design, titration, sequencing and computational analysis to reach higher level of coverage in human degraded samples. Extended downstream analysis for is required to fully evaluate this method.

Keywords: ancient DNA, next-generation sequencing, target enrichment, transposable elements, population genomics, SNP discovery.

CONTENT

1. INTRODUCTION

1. 1. Ancient DNA (aDNA) and the challenges in aDNA research

Ancient DNA (aDNA) research focuses on the isolation and subsequent analysis of DNA from archaeological specimens. The biomolecules preserved in such material can be highly degraded, but the field has recently thrived on rapid optimizations and advances. Examples include the analysis of DNA recovered from archaeological skeletal material, mummified tissues, archival collections of non-frozen medical specimens, preserved plant residues, and so on.

Compared to modern DNA, aDNA has additional challenges due to the poor quality of endogenous DNA preservation. Post mortem, taphonomic processes begin the breakdown of the body and subsequent biomolecules. This process leads to the denaturation of DNA molecules. Other factors such as soil environment and temperature can accelerate these affects, decreasing the quantity and quality of endogenous DNA.

The main issues of working with aDNA are:

1) aDNA post mortem damage, which is caused by depurination, deamination and oxidation of the nitrogen bases (Pääbo 1989; Pääbo et al. 2004). Post-mortem DNA damage begins immediately after the death of an organism. The DNA is rapidly degraded by enzymes, bacteria and fungi. UV radiation, also, produces crosslinks that will inhibit PCR (Graham 2007). The most common form of the hydrolysis is the loss of amino groups from the bases adenine, cytosine, 5-methylcytosine, and guanine, resulting in hypoxanthine, uracil, thymine, and xanthine, respectively. This causes incorrect bases (A instead of G, and C instead of T) to be inserted when new DNA strands are synthesized by a DNA polymerase. The transitions of bases C to T is presented at both ends of the molecules, but increased at the 5′-most nucleotide position in single-stranded overhangs, and G to A at the 3′-most position of molecules (Briggs et al. 2007; Jónsson et al. 2013).
2) Modern DNA contamination. Ancient DNA can be dominated by microbial DNA, which often contributes to 99% or more of the sequences. The principal modern contaminants of ancient samples are fungi and bacteria derived from

the environment where the organisms have been deposited (Willerslev and Cooper 2005). In addition, the present-day human contamination from another organism that can occur during the excavation (Gansauge and Meyer 2014), as well as airborne contaminants from the laboratory and contaminants present in laboratory reagents or on consumable items (Graham 2007).

Studies have shown that petrous bone and teeth roots are currently recognized as the optimal substrates for such research, owing to high levels of endogenous DNA (Hansen et al. 2017; Margaryan et al. 2018). In addition, new methods enable the prevention of potential contamination of ancient samples, including extraction, library and amplification controls, and the performance of all the pre-PCR steps in specialized aDNA facilities, where no DNA has been amplified or modern DNA extractions have been present. Also, with the development of Next Generation Sequencing (NGS) technologies and bioinformatics tools it is possible to evaluate the authenticity of the generated sequences and contamination levels (e.g. Skoglund et al. 2014; Jónsson et al. 2013).

1. 2. High- throughput DNA sequencing

Next-generation sequencing (NGS) refers to all the non-Sanger sequencing instruments, which are capable of producing massive amounts of parallel sequencing. Compared to Sanger sequencing that is characterized for the sequencing of specific PCR products, NGS is defined by its non-specificity. NGS platforms (e.g. Illumina MiSeq or Illumina HiSeq) use as template sequencing libraries, which are the DNA fragments from an extract that have been ligated to universal sequencing-adapters (Briggs and Heyn 2012). This application has dramatically changed the field of aDNA by enabling the retrieval of nuclear genomes from archaeological samples. It allows sequencing of the very short molecules which are characteristic of aDNA and which are generally too short to be amplified by the polymerase chain reaction (PCR). NGS thereby increases the number of endogenous ancient molecules accessible for sequencing and reduces the risk of favoring long molecules originating from modern contaminants (Knapp et al. 2012).

Importantly, the data derived from NGS can also be used to mitigate some of the challenges derived from aDNA post-mortem damage and contamination. NGS data enables us to identify and quantify the chemical and structural properties

characteristics of aDNA molecules; for instance, by estimating the average size of the sequenced material or DNA damage patterns, such as the increased frequency of C to T transition towards the 5' end of the sequencing reads (Figure 1. 1) (Ginolhac et al. 2012; Jónsson et al. 2013). This information can be used to authenticate the ancient endogenous origin of the sequenced DNA. NGS data can also be used to estimate the degree of modern contaminants. There are several methods for this, and they all rely on the principle that the sequenced DNA should only derive from one source (e.g. Skoglund et al. 2014; Renaud et al. 2015).

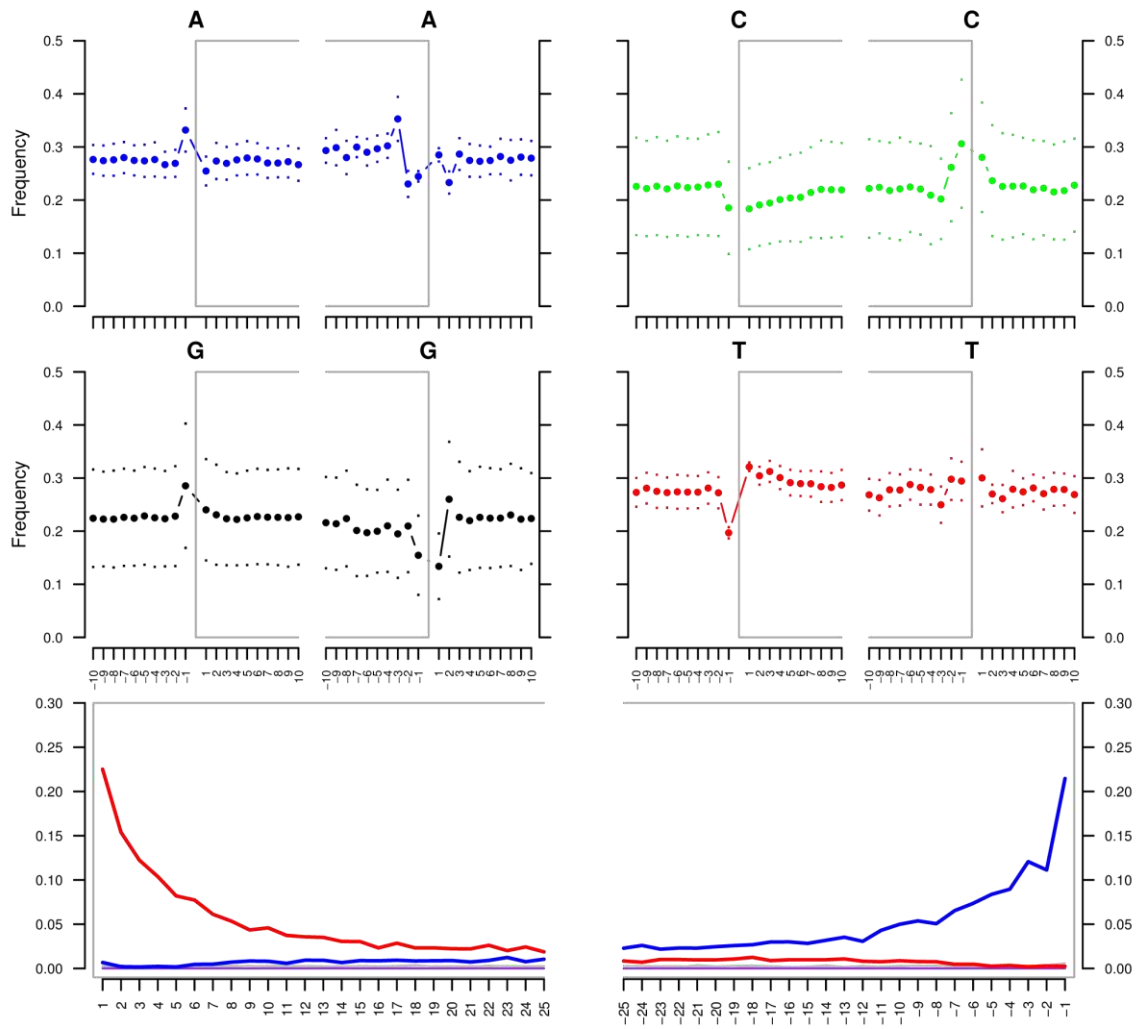


Figure 1. 1. Deamination patterns obtained using mapDamage at the end of the sequencing read. Shown on the left is the C to T deamination rate at the 5' end, and on the right the G to A rate at the 3' end. Colors correspond to three sequenced specimens. https://ginolhac.github.io/mapDamage/

Shotgun sequencing is the non-specific sequencing of all the fragments present in an extract. In aDNA, due to the limited amount of endogenous sequences compared to modern DNA contaminants, this sequencing strategy will yield a large percentage of sequences that will be discarded (e.g. fungi and bacteria) (Knapp and Hofreiter 2010). Usually in aDNA samples the fraction of authentic DNA is represented by less than 1% of the total DNA present in the sample, so recovering entire genomes requires an increased amount of costly sequencing. As a result, efforts are constantly being made to optimize access to the endogenous DNA of ancient samples and to improve the recovery of aDNA sequence data.

There are two main areas in which researchers in the aDNA community have focused on in the past for improving the access to authentic, endogenous aDNA:

1) optimizing existing methods for DNA extraction and library preparation for ancient samples (Rohland and Hofreiter 2007a, 2007b; Adler et al. 2011; Damgaard et al. 2015; Glocke and Meyer 2017; Carøe et al. 2018; Rohland et al. 2018) and

2) devising new ways for the targeted retrieval of aDNA sequencing data through hybridization capture and related approaches (Carpenter et al. 2013; Ávila-Arcos et al. 2015; Cruz-Dávalos et al. 2017b).

2.STATE OF THE ART

2. 1. Reduced representation library (RRL) methods

Population genomics refers to the simultaneous study of numerous loci or genome regions to understand the roles of evolutionary processes (e.g. mutation, random genetic drift, gene flow or natural selection) that create variation across genomes and populations (Luikart et al. 2010). The development of population genomics occurred together with the advances in NGS technologies and methods. Even though the most robust and reliable population genomic inferences are derived from whole genomes and resequencing experiments (Luikart et al. 2003; Fuentes-Pardo and Ruzzante 2017), these are still very expensive for most research institutions (Andrews et al. 2016).

Reduced representation genomic libraries (RRLs) are increasingly used to reveal diversity in evolutionary biology; Reduced representation library sequencing approaches have been developed to select a subset of the genome, reducing the cost of sequencing. Thus, reduced representation library (RRL) methods have become a popular alternative for SNP discovery and genotyping (Davey et al. 2011), in particular for non-model organisms. Several RRL strategies have been developed in the last years, including restriction site-associated DNA sequencing (RADseq) (Baird et al. 2008; Davey and Blaxter 2010), double digest RADseq (ddRADseq) (Peterson et al. 2012) or genotyping-by-sequencing (GBS) (Elshire et al. 2011), as well as the combination of RRLs with hybridization by capture for genotyping museum and ancient specimens (Barreiro et al. 2017; Schmid et al. 2017).

All these methods combine the use of restriction enzymes to cut DNA into fragments and the transformation of the sheared material into sequencing libraries (usually Illumina) (Andrews et al. 2016; Davey and Blaxter 2010). Sequencing libraries will present barcoded-adapters, so several individuals (up to hundreds) can be pooled and sequenced together, and then bioinformatically assigned the sequencing reads to particular individuals. All these methods start with relatively high molecular weight genomic DNA (C. F. Graham, Glenn, and McArthur 2015) and begin by digesting it with one (e.g. RADSeq) or more (e.g. GBS) restriction enzymes. After DNA shearing by restriction enzymes, specific sequencing adapters, that are required by NGS

platforms, are ligated to the fragmented DNA. Their advantages over whole-genome sequencing are:

I. great depth of coverage per locus, which will improves confidence in genotype calls, and
II. cost reduction, that will allow researchers to study higher number of samples (Andrews et al. 2016).

2. 1. 1. RADSeq methods

In RADSeq (Figure 2. 1) fragments are ligated to P1 adapters (sample specific barcodes), pooled and size selected to 300–700 bp (Baird et al. 2008). Y-shaped P2 universal sequencing adapters (Coyne et al. 2004) are ligated to the fragments with and without P1 adapters. Prior to sequencing, fragments will be PCR amplified with P1 and P2 specific primers. Thus, only fragments with both P1 and P2 adapters will be amplified (i.e. only those sites that have been digested with the restriction enzyme).

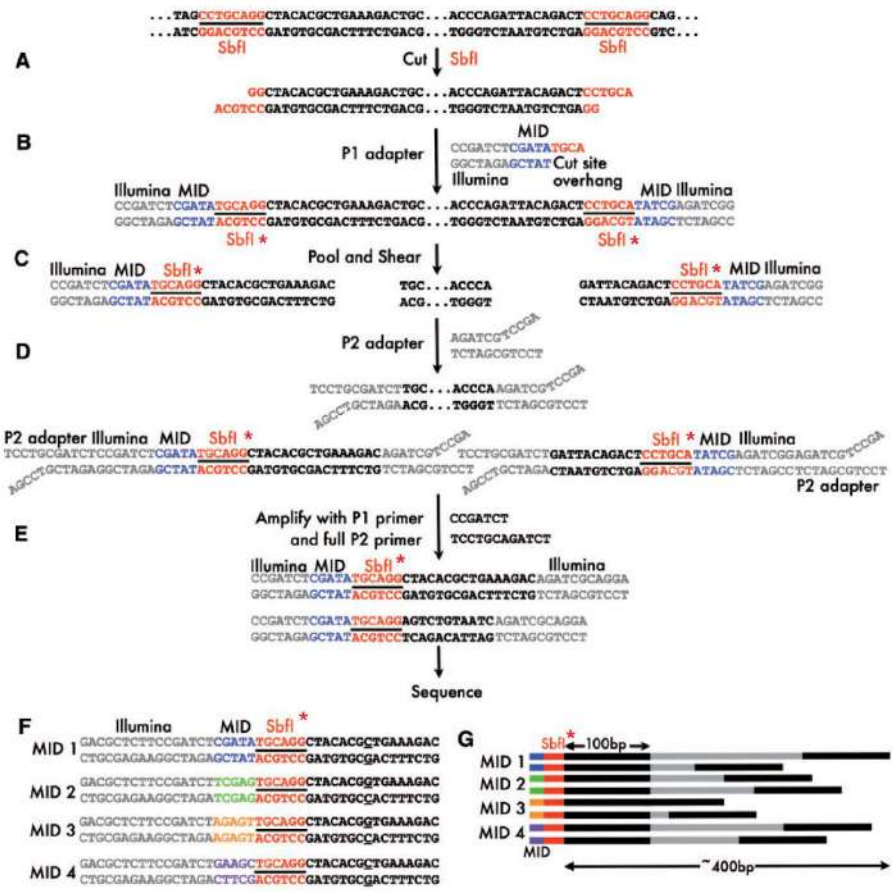


Figure 2. 1. The process of RADSeq.

(A) Genomic DNA is sheared with a restriction enzyme of choice (SbfI in this example). (B) P1 adapter is ligated to SbfI remained (SbfI*)-cut fragments. The P1 adapter is adapted from the Illumina sequencing adapter (full sequence not shown here), with a molecular identifier (MID; CGATA in this example) and a cut site overhang at the end (TGCA in this example). (C) Samples from multiple individuals are pooled together and all fragments are randomly sheared. Only a subset of the resulting fragments contains restriction sites and P1 adapters. (D) P2 adapter is ligated to all fragments. The P2 adapter has a divergent end. (E) PCR amplification with P1 and P2 primers. The P2 adapter will be completed only in the fragments ligated with P1 adapter, and so only these fragments will be fully amplified. (F) Pooled samples with different MIDs are separated bioinformatically and SNPs called (C/G SNP underlined). (G) As fragments are sheared randomly, paired end sequences from each sequenced fragment will cover a 300- 400 bp region downstream of the restriction site (Davey and Blaxter 2010).

RADseq opens up rich prospects for analysis of genetic markers, both in the detailed information that can be gained from single markers and from the complex interactions between thousands of markers across the genome.

2. 1. 2. Double digest RADseq (ddRADseq) method

Double digest RADseq (ddRADseq) method eliminates random shearing and end repair of genomic DNA using a double restriction enzyme (RE) digest (i.e., a restriction digest with two enzymes simultaneously). The selection for genomic fragments by size allows greater fine-scale control of the fraction of regions represented in the final library. By combining precise and repeatable size selection with sequence-specific fragmentation, ddRADseq produces sequencing libraries consisting of only the subset of genomic restriction digest fragments generated by cuts with both REs (i.e., have one end from each cut) and which fall within the size-selection window (Figure 2. 2).

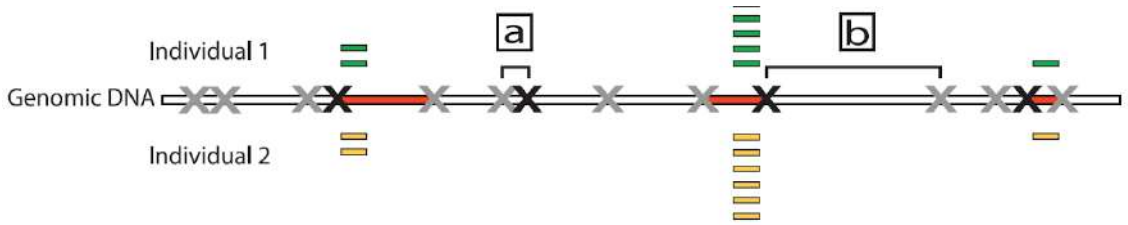


Figure 2. 2. The process of double digest RAD sequencing.
Double digest RAD sequencing (ddRADseq), by contrast, uses a two enzyme double digest followed by precise size selection that excludes regions flanked by either [a] very close or [b] very distant RE recognition sites, recovering a library consisting of only fragments close to the target size (red segments). Representation in this library is expected to be inversely proportional to deviation from the size-selection target, thus read counts across regions are expected to be correlated between individuals (yellow and green bars) (Peterson et al. 2012).

This combination of requirements can be tuned to generate libraries consisting of fragments derived from hundreds to hundreds of thousands of regions genome-wide due to the removal of random shearing (and therefore random recovery), correlated recovery of regions across individuals results in increased robustness to variability in read count. As sequencing depth required to reach saturation, the number of

individuals which can be genotyped in a single sequencing lane is inversely proportional to the number of regions recovered.

2. 1. 3. Genotyping by Synthesis (GBG) method

GBS protocol (Figure 2. 3) requires less laboratory steps compared to RADSeq. In GBS, barcoded adapters (green and yellow) and universal adapters (grey) are ligated to digested fragments (Elshire et al. 2011). As a result, there will be a DNA mixture of fragments with barcode+universal, barcode+barcode and universal+universal adapter combinations (Davey et al. 2011). Samples are pooled and subsequently amplified on the Illumina Genome Analyzer flow cell. Only samples featuring a barcode+common adapter combination will be amplified for sequencing.

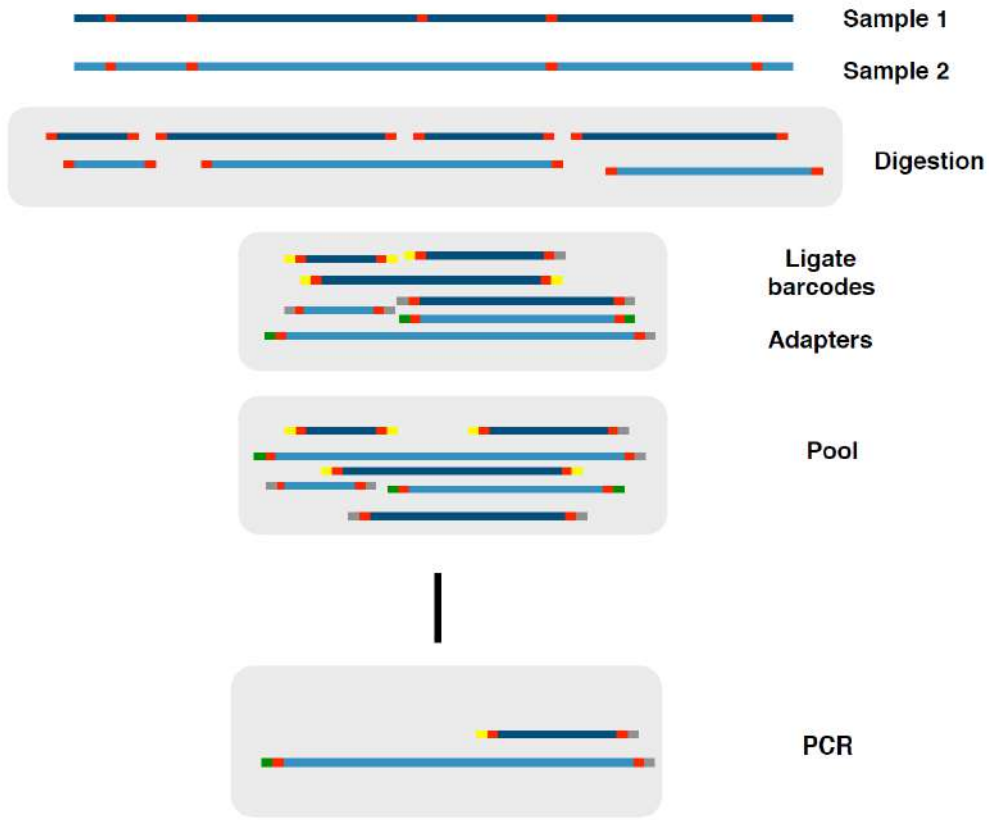


Figure 2 .3. GBS protocol, adapted from Davey et al. (2013).

2. 1. 4. Challenges in reduced representation Library (RRL) sequencing methods

Despite the advantages that RRL methods present, there are also challenges associated with them. First of all, the required DNA quality in the starting material;

RRL methods require high molecular weight DNA for enzymatic digestion, which makes them not applicable to degraded samples (e.g. museum specimens) (Davey et al. 2011). RADseq requires especially high molecular weight compared to other methods, as the consistency and efficiency of the mechanical shearing step would depend on relatively large size (Andrews et al. 2016).

Another issue is the amount of starting material recommended for these methods. In general, large amounts of DNA are preferred, as this will reduce the number of PCR cycles during the protocol and, as a consequence, the number of sequenced PCR duplicates. The first version the RADSeq protocol required up to 1 µg per sample (Etter et al. 2011). Currently, most of these protocols can often be implemented with only 50–100 ng of DNA per sample (Andrews et al. 2016).

One of the main challenges of RRL methodologies is associated with allele dropout and null alleles. RRL datasets are likely to present high proportions of missing data, mostly due to mutations in the restriction enzyme recognition site (Gautier et al. 2013). Mutations in the enzyme restriction site will result in the failure to cut the genomic DNA at that location (i.e. allele dropout). Null alleles will derive from alleles that lack the recognition site and, thus, they will not be sequenced (Huang and Knowles 2014). The effect of allele dropout and null alleles will be that even though a high number of loci are sequenced per sample, the number of comparable sites can become highly reduced (Gautier et al. 2013), and if a single nucleotide polymorphism (SNP) occurs within a null allele, this could derive in genotyping errors (i.e. heterozygous individuals for the null allele appearing as homozygotes) (Andrews et al. 2016).

Clonal DNA fragments might be generated during these PCR steps, which are known as PCR duplicates (Davey et al. 2011). PCR duplicates need to be identified and filtered out during data processing, as they would inflate coverage estimates, cause a heterozygote to look like a homozygote during genotype calling, or make an allele containing a PCR error appear to be an actual allele (false allele) (Andrews and Luikart 2014). Several studies have reported that PCR duplicates occur at high frequencies in RADseq data (e.g. Andrews et al. 2014; Schweyen et al. 2014). Clonal sequences will also be taking up sequencing capacity that could have been allocated to generate non-clonal reads. PCR duplicates can be identified in RAD

protocols that include a random shearing step and paired-end sequencing, like the original RADseq (Davey et al. 2013; Andrews and Luikart 2014). However, PCR duplicates cannot be identified in some other RRL strategies, because all fragments for a given locus will have identical start and stop positions (Davey et al. 2011; Andrews et al. 2016). Alternatives for controlling PCR duplicates are using PCR-free protocols, such as ezRAD that relies on Illumina PCR-free kits for library build. However, PCR free methods are frequently more expensive and require larger amounts of starting material (Andrews et al. 2016).

Other aspect that needs to be taken into consideration in RRL experiments is the variance in depth of coverage among loci, which would require an increase in sequencing effort in order to achieve similar values across loci (Andrews et al. 2016). General recommendation is to remove those loci with inconsistent coverage, as they could lead to genotyping biases and thus influence population inferences (Andrews and Luikart 2014). G-C content biases will also affect those RRL protocols that include PCR steps by influencing depth of coverage.

Despite all the challenges associated with RRL methodologies, they are powerful and versatility tools for SNP discovery and genotyping in ecological and evolutionary genomics, such as population genomic (e.g. (Hohenlohe et al. 2010)), phylogeographic (e.g. Gaither et al. 2015; Emerson et al. 2010), and phylogenomic (e.g. Wagner et al. 2013) studies. In the last few years, hybridization capture strategies have been implemented to RRL sequencing. These methods rely on the synthesis into capture baits of the RRL loci (e.g. Ali et al. 2016; Schmid et al. 2017). Capture of the RRL loci prior to sequencing allows the generation of RRL genome-wide data in degraded samples. Another advantage of hybridization capture of RRL loci is the reduction in allele dropout. However, bait design can be complex and costly, as some of these protocols require commercially synthesized probes (Schmid et al. 2017).

2. 2. Enrichment methods

The application and development of enrichment by capture methods have been of special interest for the aDNA community. In addition, massively parallel DNA sequencing technologies have increased the ability to generate large amounts of sequencing data at a rapid pace. Several methods have been developed to enrich

for genomic regions of interest for sequencing using a combination of target capture enrichment methods and NGS technologies. Although, cost and technology are still limiting factors in aDNA sequence.

2. 2. 1. Target capture methods

Targeted genome capture (TGC) is a methodology that enriches specific genetic sequences within a heterogeneous mixture of DNA or RNA (Gnirke et al. 2009). In targeting capture methods molecular markers can be used to obtain extensively genetic information of high-interest genes. The definition of “marker” in DNA is any informative region of the genome that is either non-protein coding and therefore selectively neutral, or is protein coding and therefore potentially subject to selection (Cabana, Hulsey, and Pack 2013). Different approaches to generate enriched fragments of targeted DNA have been developed the past years (Figure 2. 4).

First, PCR has been the most widely used pre-sequencing sample preparation technique for over 20 years (Saiki et al. 1988). PCR is potentially compatible with any next generation sequencing platform, though to make full use of the high throughput, a large number of amplicons must be sequenced together. However, PCR is difficult to multiplex to any useful degree: the simultaneous use of many primer pairs can generate a high level of nonspecific amplification, caused by interaction between the primers. Moreover amplicons can fail to amplify sufficiently (Cho et al. 1999; Wang 1998).

Second, molecular inversion probes (MIPs) had been developed for multiplex target detection and SNP genotyping (Hardenbol et al. 2005; Lizardi et al. 1998). A newly developed MIP capture method is based on the following characteristics:

I. gap-fill reactions and PCRs take place in aqueous solution, in small volumes
II. sample-identifying barcodes are nested in one of the primers used in post-capture amplification, allowing products from multiple samples to be pooled and sequenced in a single lane
III. as with PCR, capture is performed directly on genomic DNA rather than after conversion to a shotgun library, reducing input requirements (Deng et al. 2009).

Although, MIP oligonucleotides can be costly and difficult to obtain in large numbers to cover large target sets.

Finally, in the hybridization capture the principle of direct selection is well-established (Lovett et al. 1991; Parimoo et al. 1991): a shotgun fragmented library is hybridized to an immobilized probe, nonspecific hybrids are removed by washing and the targeted DNA is eluted. More detailed, the array-based hybridization method is based on high-density microarrays containing probes complementary to the regions of interest to bind and purify DNA molecules of interest. Microarrays are glass slides densely spotted with clusters of single-stranded synthetic oligonucleotides that are allowed to hybridize with fluorophore-labelled DNA from a sample, and the resulting fluorescence signals are interpreted to determine sequence composition and/or taxonomic content (Devault et al. 2014). The second approach of hybridization capture is a solution-based method that uses biotinylated DNA or RNA complementary probes to bind to targets. The latter approach has several advantages over the array-based methods; as it is a highly scalable technique that does not require additional equipment associated with processing microarrays (Bodi et al. 2013).

There are several factors affecting capture assays, such as starting material, probe tiling, hybridization temperature, or the proportion of endogenous DNA. Additionally, when designing the capture assay, many parameters should be considered in evaluating the performance of each approach.
Those factors are:

I. sensitivity, or the percentage of the target bases that are represented by one or more sequence reads
II. specificity, or the percentage of sequences that map to the intended targets
III. uniformity, or the variability in sequence coverage across target regions
IV. reproducibility, or how closely results obtained from replicate experiments correlate
V. cost
VI. ease of use and
VII. amount of DNA required per experiment, or per megabase of target.

Researchers that want to perform enrichment by capture can choose between do-it-yourself protocols by which DNA or RNA baits will be generated in-house (e.g. Carpenter et al. 2013; Maricic et al. 2010), or ordering commercial generic or

customized capture kits (e.g. MYbaits from MYcroarray). In this way, a large number of experimental parameters may affect the efficacy of aDNA target enrichment using DNA probes. The annealing temperature is a crucial factor over all capture stringency, with high temperatures (65°C) leading to increased specificity and coverage (Cruz-Dávalos et al. 2017; Paijmans et al. 2016).

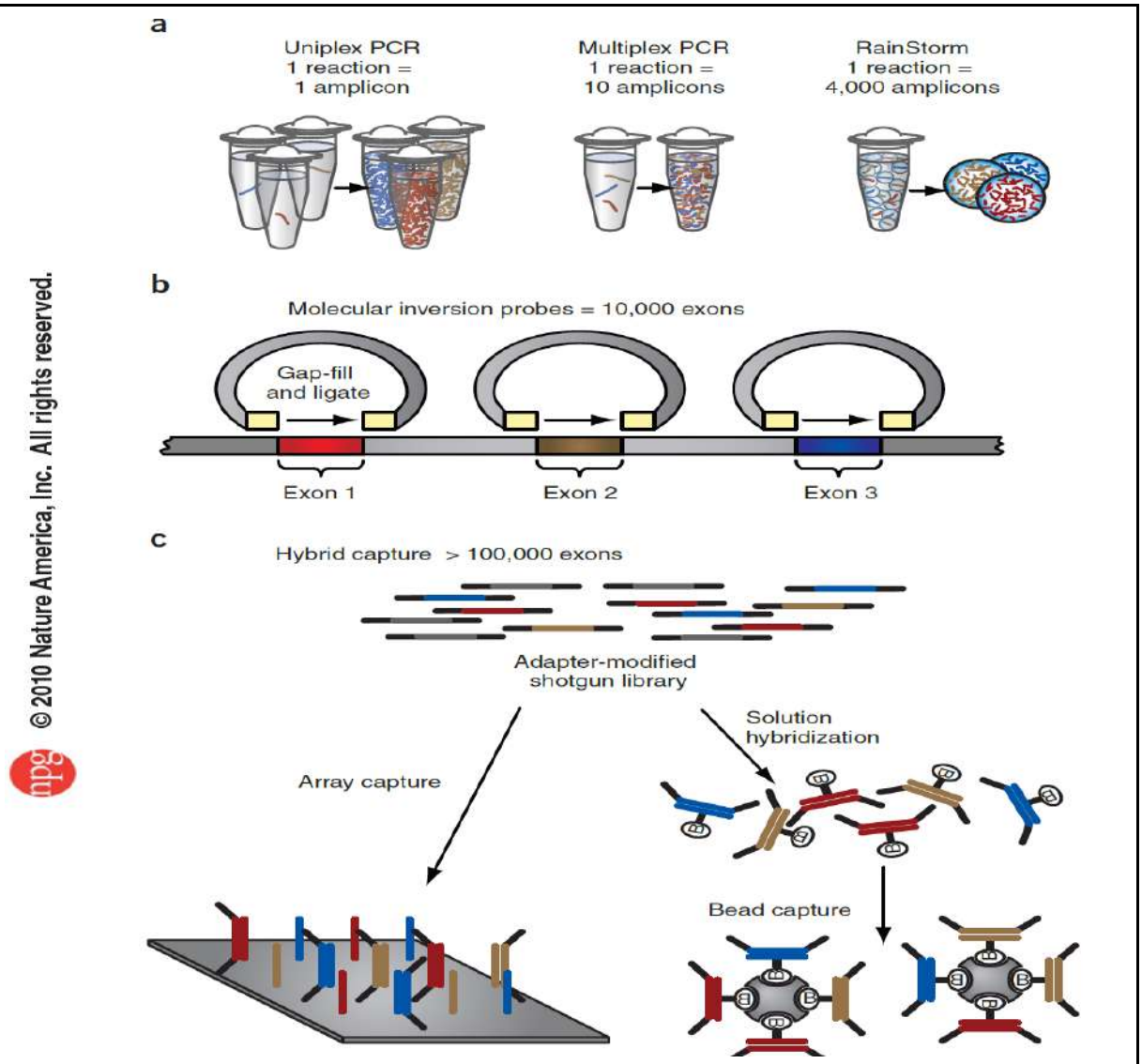


Figure 2. 4. Approaches to target enrichment.

(a) In the uniplex PCR–based approach, single amplicons are generated in each reaction. In multiplexed PCR, several primer pairs are used in a single reaction, generating multiple amplicons. (b) In the MIP-based approach, probes consisting of a universal spacer region flanked by target-specific sequences are designed for each amplicon. These probes anneal at either side of the target region, and the gap is filled by a DNA polymerase and ligase. Genomic DNA is digested, and the target DNA is PCR-amplified and sequenced. (c) In the hybrid capture–based approach, adaptor modified genomic DNA libraries are hybridized to target-specific probes either on a microarray surface or in solution. Background DNA is washed away, and the target DNA is eluted and sequenced (Mamanova et al. 2010).

2. 2. 1. 1. Mitochondrial DNA (mtDNA) capture methods

Mitochondrial DNA (mtDNA) is a separate genome located in the cytoplasm of nearly all eukaryotic cells (Anderson et al. 1981) and it is also compact, circular, and double-stranded (Gray 2001).

Capturing the mitochondrial DNA opened new possibilities in ancient DNA research, especially in contexts where destructive analysis of skeletal material is not possible. Different research group achieved to enrich complete mitochondrial genomes from complex samples, such as saliva, dental calculus and ancient hominin bones, using mtDNA capture methods (Ozga et al. 2016; Maricic et al. 2010).

Two different methods have been developed the past years for mtDNA enrichment. The primer-extension-capture (PEC) method (Fig. 2. 5) and long range PCR method (Fig. 2. 6) are used to capture mtDNA (Ozga et al. 2016; Maricic et al. 2010; Briggs et al. 2009).

The primer extension capture (PEC) method directly isolates specific DNA sequences from complex libraries of highly degraded DNA. PEC uses 5'-biotinylated oligonucleotide primers and a DNA polymerase to capture specific target sequences from an adaptor-ligated DNA library. The high specificity of PCR primers gives the advantage of the immortalization through reamplification from adaptor priming sites (Blow et al. 2008), contamination control with project-specific barcodes (Briggs et al. 2007; Green et al. 2008), access to very short fragments predominant in ancient extracts (Brotherton et al. 2007), and quantification of the number of unique ancient DNA molecules, which is necessary to identify nucleotide misincorporations (Brotherton et al. 2007; Mackelprang and Rubin 2008; Briggs et al. 2007).

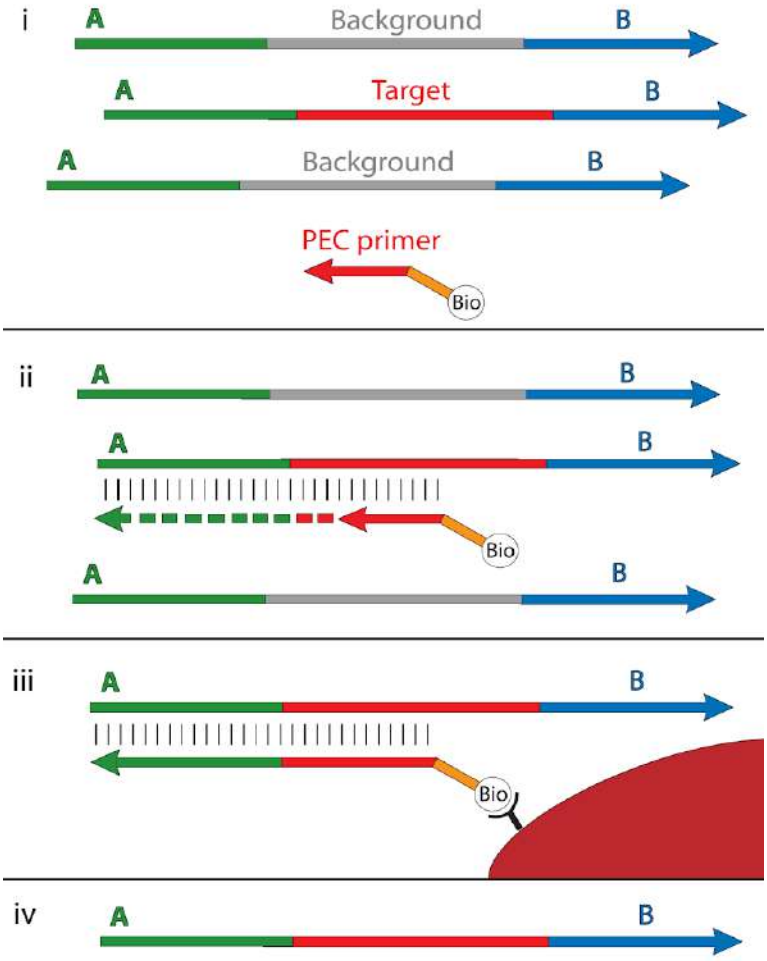


Figure 2. 5. Primer extension capture (PEC).

(i) 5′-Biotinylated oligonucleotide primers (PEC primers) are added to a 454 library [in which the A and B adaptor molecules carry a project-specific barcode] and are allowed to anneal to their respective target sequences. (ii) A single Taq DNA polymerase extension step is performed, resulting in a double-stranded association between primer and target that includes the 5′ adaptor sequence. (iii) Excess PEC primers are removed by spin column purification, and the biotinylated primer:target duplexes are captured by streptavidin- coated magnetic beads. The beads are washed stringently above the melting temperature of the PEC primers, to ensure that templates upon which extension occurred will preferentially remain associated with the primers. (iv) Captured and washed targets are eluted from the beads, amplified with adaptor priming sites, and subjected either to a second round of extension and capture or directly to 454 emulsion PCR (after Briggs et al. 2009).

Methods that are able to capture relevant DNA sequences rely on hybridization of target sequences to probes that can be either in solution or immobilized on a surface, as previously described. The long range PCR method, uses the PCR products to capture targets for sequencing from pooled libraries of multiple individuals, using standard laboratory equipment. DNA baits for complete human mitochondrial genome were usually produced from modern DNA extracts using previously described primers (Meyer et al. 2007; Ozga et al. 2016). This method is applied to DNA pools of libraries from several human individuals from which are captured complete mtDNAs, which are extensively studied in population genetics, medicine, forensics, and phylogenetics (Pakendorf and Stoneking 2005).

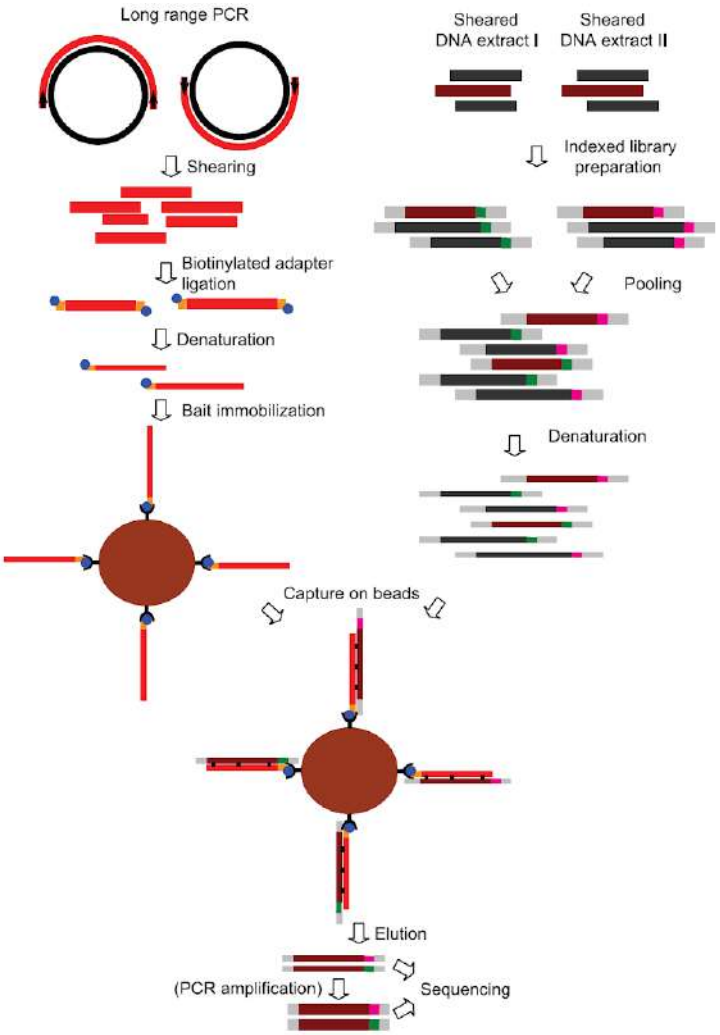


Figure 2. 6. Overview of the capture-on-beads method.

On the left the production of the immobilized bait from two long range PCR products is shown; on the right the production of a pool of indexed libraries which are used in the capture (bottom). The eluted molecules can either be sequenced directly or first amplified and then sequenced. The bait is light red, mitochondrial DNA in the libraries is dark red, indices are shown in green and pink, adapters in gray. Thicker lines represent double stranded DNA while thinner lines represent single stranded DNA. (after Maricic, Whitten, and Pääbo 2010).

Improvements in mtDNA capture enrichment efficiency are necessary, but even current methods are sufficient for full mitogenome reconstruction from small quantities. MtDNA enrichment is cost efficient, because it requires only standard laboratory equipment and reagents, and fast. Additionally, many approaches can be

multiplexed, allowing efficient analysis of many samples in parallel. Although, these novel methodologies can be only applied in mitochondrial DNA.

2. 2. 1. 2. Chromosomal capture methods

The ancient Y chromosome sequences are providing the first glance into the past variation of male specific compartment of the genome and the opportunity to evaluate models based on previously made inferences from patterns of genetic variation in living populations. Ancient Y chromosomes allows also a better understanding of the rate at which mutations accumulate and get fixed over time.

Analyses of the ancient Y chromosome sequences are challenging not only because of issues generally related to ancient DNA work, such as DNA damage-induced mutations and low content of endogenous DNA in most human remains, but also because of specific properties of the Y chromosome, such as its highly repetitive nature and high homology with the X chromosome (Kivisild 2017).

Previous studies (Burbano et al. 2012; Avila-Arcos et al. 2011; Burbano et al. 2010) have shown that hybridization enrichment can be used to obtain nuclear DNA fragments from ancient samples. Chromosomal capture was based on oligonucleotides synthesized on arrays to construct probe libraries that are amplified and converted into biotinylated DNA/ RNA capture probes through in vitro transcription. The probes were designed using the human reference genome sequence (hg19) (Fu et al. 2013).

Summarizing, a number of aDNA studies have already started to reveal the potential of human Y chromosome to inform about the demographic past, but shotgun sequencing of uniquely mapping regions of the Y chromosomes with sufficiently high coverage is still challenging and costly in degraded samples.

2. 2. 1. 3. Pathogen capture methods

Pathogen capture materialized by array- based capture screening technique. These methods utilize oligonucleotide probes to enrich specific nucleic acids in heterogeneous extracts and can therefore increase the proportion of NGS reads for low-abundance targets. Two different pathogen capture methods have developed the past years. Both methods use arrays to target specific pathogen sequences.

The first method, called ancient pathogen screening array (APSA), combines the DNA capture coupled with next generation sequencing for parallel detection of ancient pathogens (Figure 2. 7).

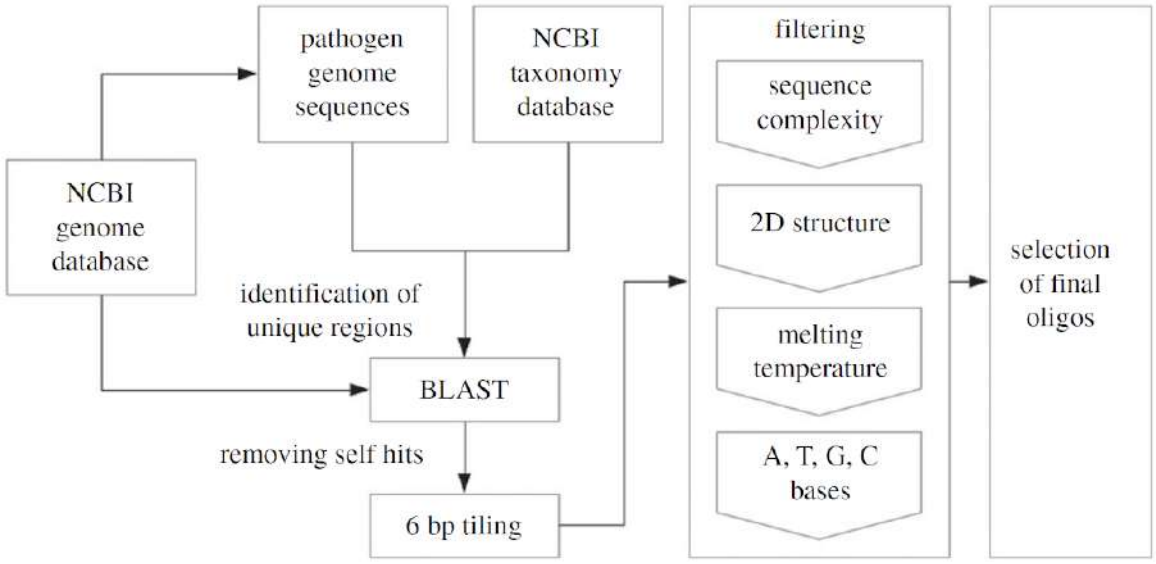


Figure 2. 7. Workflow followed in probe design APSA.

Selection of capture regions was performed based on NCBI taxonomic relationships (Federhen 2012). Candidate regions were then screened for uniqueness by BLAST searches against the NCBI nucleotide database. Probes 60 bp in length were generated at 6 bp tiling for each genomic region considered suitable for capture design. Oligos were subsequently filtered. The final selected probes were permitted on an Agilent 1-million feature array, and efforts were made to provide equal representation of all pathogens, regardless of genome size (Bos et al. 2014).

Shotgun metagenomics offers a powerful tool to fully characterize pathogens from ancient samples. However, High-throughput sequencing (HTS) is only useful when the primary pathogen(s) are known or suspected to be present, and ignores non-targeted taxa and genomic loci. HTS metagenomic approaches can be labour- and time-intensive, thereby representing significant barriers for groups that would like to thoroughly profile or screen the microbial content of large or difficult paleopathological sample sets.

The second recently developed method for pathogen capture uses microbial detection array. The Lawrence Livermore Microbial Detection Array (LLMDA) (Gardner et al. 2010) contains probes designed from all published vertebrate-infecting pathogen genomes. LLMDA probes target conserved regions amongst all known species/strains of a family (or equivalent unit), but due to the high number and overall diversity of probes, unique combinations of matching probes across an individual genome sequence allow for species or strain identification. Fluorescence data are analysed using a likelihood maximization algorithm to identify the combination of species that best explains the resulting signal. To achieve this, each signal set is compared against a current database of full microbial genomes and analysed for the expected vs. detected combined probe fluorescence signal, resulting in a species list ranked by likelihood of presence (Devault et al. 2014).

This method provides similar bacterial family-level metagenomic profiles of archaeological and archival specimens as HTS, especially for the most abundant taxa, and successfully detected the previously-verified infecting pathogen species in both specimens, but it needs progress to become an excellent screening tool for archaeological samples where microbial profiles can be swiftly, cheaply, and accurately reconstructed thereby aiding the elucidation of population health through deep time.

2. 2. 2. Whole genome capture method

Most of the existing WGC methods for ancient DNA require expensive laboratory equipment and protocols. A method referred to as whole genome in-solution capture (WISC) succeeds to increase the proportion of endogenous DNA in aDNA sequencing libraries (Figure 2. 8). WISC uses in vitro transcription of DNA libraries with biotinylated UTP, producing RNA baits covering the entire human genome. Analogous to current exome capture technologies (Gnirke et al. 2009) these baits are hybridized to aDNA libraries in solution and pulled down with magnetic streptavidin-coated beads.

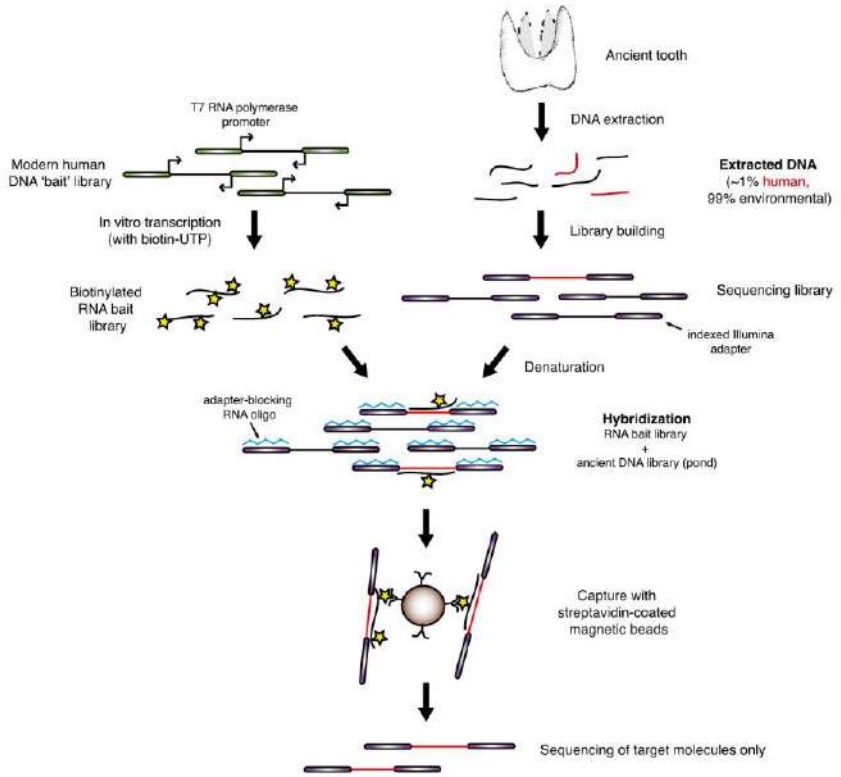


Figure 2. 8. Schematic overview of the Whole-Genome In-Solution Capture Process.

To generate the RNA ‘‘bait’’ library, a human genomic library is created via adapters containing T7 RNA polymerase promoters (green boxes). This library is subjected to in vitro transcription via T7 RNA polymerase and biotin-16-UTP (stars), creating a biotinylated bait library. Meanwhile, the ancient DNA library (aDNA ‘‘pond’’) is prepared via standard indexed Illumina adapters (purple boxes). These aDNA libraries often contain <1% endogenous DNA, with the remainder being environmental in origin. During hybridization, the bait and pond are combined in the presence of adaptor-blocking RNA oligos (blue zigzags), which are complementary to the indexed Illumina adapters and thus prevent nonspecific hybridization between adapters in the aDNA library. After hybridization, the biotinylated bait and bound aDNA is pulled down with streptavidin-coated magnetic beads, and any unbound DNA is washed away. Finally, the DNA is eluted and amplified for sequencing (Carpenter et al. 2013).

WISC can be used to highly enrich the endogenous contents of aDNA sequencing libraries, thus reducing the amount of sequencing required to sample the majority of

unique fragments in the library with low cost, but the recovery of endogenous content can be really poor (~1%).

2. 2. 3. 1240k capture method

The crucial problem of the recovery of aDNA data is that it is typically very low coverage. A small number of samples have high endogenous content and complexity, therefore can be sequenced to high coverage using 1240k targeted capture enrichment approach.

The 1240k method is an in-solution hybridization capture technique with synthesized oligonucleotide probes to enrich promising libraries for more than 1.2 million SNPs ("1240k capture", Methods). The targeted sites include nearly all SNPs on the Affymetrix Human Origins and Illumina 610-Quad arrays, 49,711 SNPs on chromosome X and 32,681 on chromosome Y, and 47,384 SNPs with evidence of functional importance. It merge libraries from the same individual and filter out samples with low coverage or evidence of contamination to obtain the final set of individuals (Mathieson et al. 2015).

In conclusion, the advantage of the 1240k capture approach is that it accessed to genome-wide data from ancient samples with small fractions of human DNA and increases efficiency by targeting sites in the human genome that will actually be analyzed. Although, the limitations of this method are the high cost and restricted availability.

2. 3. Mobile elements Sequencing (MobiSeq) RLL protocol

Hybridization capture of reduced representation library (RRL) loci allows the sequencing of RRL loci in degraded samples and reducing allele dropout. MobiSeq is a RRL protocol exploiting simple laboratory techniques, that generates genomic data based on PCR targeted-enrichment of transposable elements (TEs) and the sequencing of the associated flanking region. TEs are self-replicating mobile elements that insert themselves in new places of the genome, either through a cut-and-paste or a copy-and-paste mechanism (Kazazian 2004). The method has been previously tested on modern samples (Rey-Iglesia, Gopalakrishnan, and Carøe 2018) and enables the sequencing of hundreds of thousands loci across the genome, and performs SNP discovery with relatively low rates of clonality in modern

samples. Given the ease and flexibility of the MobiSeq protocol, the method has the potential to be successful in ancient samples, too. However, given the fragmented nature of ancient DNA, it remains to be seen to what extent it can be successfully applied.

2. 3. 1. Transposable genetic elements (TEs)

In 1976, Bukhari et al. (1976) defined transposable elements (TE) as DNA sequences that could move. TEs are repetitive and mobile DNA sequences, with the ability to integrate into the genome at a new site within their cell of origin (Chénais et al. 2012; Platt et al. 2018). Transposable genetic elements (TEs) are ubiquitous in both prokaryotes and eukaryotes. TEs display little insertion preference and can be scattered throughout the genome, although they are negatively selected in exonic regions (Sela et al. 2010; Kazazian 2004).

Due to the vast contribution of TEs to genome architecture, TEs have had significant impacts on mammalian evolution (Kazazian 2004; Chénais et al. 2012; Platt et al. 2018). TEs have been reported to have a role in small-scale changes in linkage groups, as well as large structural genomic variation, such as deletions, inversions, duplications, and translocations (Gray 2000; Grabundzija et al. 2016). The mobility of TEs can induce the appearance of mutations and changes in gene expression (Chénais et al. 2012). Speciation events have been associated with the expansion of TEs in the genome (Platt, Vandewege, and Ray 2018), which suggests their potential role as drivers of adaptation, diversification, and speciation by generating structural genomic diversity between populations (Chénais et al. 2012).

TEs are classified in two major groups, (i) DNA transposons and (ii) retrotransposons (Kazazian 2004). DNA transposons move in the genome by a “cut-and-paste” mechanism, involving the excision and reinsertion of the DNA sequence of the element, or by using a rolling circle process or a virus-like process (Kazazian 2004; Chénais et al. 2012). On the other hand, retrotransposons, move in the genome via a “copy-and-paste” mechanism directed by reverse transcription of an RNA intermediate of a source element (Mourier and Willerslev 2009; Kazazian 2004).

Retrotransposable elements (or type II elements), are predominant in mammalian genomes (with bats as a notable exception) (Rey-Iglesia, Gopalakrishnan, and

Carøe 2018). Long interspersed nuclear elements (LINEs) are long retrotransposable elements encoding the enzymatic machinery required for their own movement. In contrast, short interspersed nuclear elements (SINEs) are short transcribed sequences – often derived from small RNA genes – that do not encode any proteins, instead relying on proteins encoded by LINE elements (Dewannieux, Esnault, and Heidmann 2003). Insertions of LINEs and SINEs take place through the so-called target-primed reverse transcription, ensuring that the 3' end of the elements are always present whereas the 5' may be truncated (Luan et al. 1993).

The human genome browser hosted by the University of California, Santa Cruz (Kent et al. 2002), currently contains over 4 million annotated transposon copies belonging to at least 848 families and subfamilies of elements (http://genome.ucsc.edu). These transposons collectively occupy almost half (44%) of the human genome and, thus, are major components of human genes and chromosomes (http://genome.ucsc.edu). Most of the largest transposon families in humans initially were identified as dispersed repetitive sequences that contain hallmark features of transposons (e.g. target site duplications, terminal repeats and transposases (Jurka 2000; Smit and Riggs 1996). Subfamilies are defined by specific sets of sequence changes that can be useful for tracking the evolution and activity of elements. Repbase (Jurka 2000) can be consulted for additional information on human transposons and their subfamilies (http://www.girinst.org/repbase/index.html).

Transposon subfamilies are categorized by unique sequence changes that act as markers to indicate a common phylogeny among subfamily members. Diagnostic changes could supervene in a progenitor element and copy to all new transposon insertions that are derived from this founder or its progeny.

TE-target primers can be designed to enrich any TE element present in the species of interest, making it a very flexible protocol for use on eukaryotic genomic DNA. Furthermore, several TEs can be combined, in order to increase the number of sequenced markers, thus increasing the proportion of genome coverage and analytical resolution (Rey-Iglesia, Gopalakrishnan, and Carøe 2018). Designing TE-target primers in ancient DNA is more challenging than in modern DNA. The main factors in selecting primers for ancient human DNA samples are the representation

of the transposable elements in human DNA and the mutations that accumulate over time.

2. 3. 2. Alu Elements in human DNA

Alu is the most frequent repeat and successful of all mobile elements in the human genome with more than one million copies per haploid genome (Lander et al. 2001; Chen et al. 2009) (contributing almost 11% of the human genome). Such high copy number elements could be really useful in target-enrichment methods used as “genetic marker”.

Alus appear only in the nuclear genome and are classified as retroelements termed SINEs (short interspersed elements). Alu SINEs have identified originally almost 30 years ago as a component in human DNA. They are non-autonomous and carry a Pol III promoter in their 5’. They harbor poly-A elements and CpG domains. Their ancestor seems to be the 7SL tRNA and they are flanked by short direct repeats (Rowold and Herrera 2000). Alu elements acquire trans-acting factors for their amplification from the only active family of autonomous human retroelements: LINE-1 (Dewannieux, Esnault, and Heidmann 2003).

The length of Alu elements are ~300 bp long and are commonly found in introns, 3′ untranslated regions of genes and intergenic genomic regions. They are distributed within the human genome in a defined way, as they accumulate preferentially in gene rich regions (Lander et al. 2001; Korenberg and Rykowski 1988; Chen et al. 2002). More specifically, they tend to accumulate in GC-rich regions (Jurka et al. 2007) and participate in the architecture of the genome by delimiting the active/ inactive domains and the epigenetic landscape (Edwards et al. 2010) and gene regulation at different levels (Daniel et al. 2014; Cordaux and Batzer 2009).

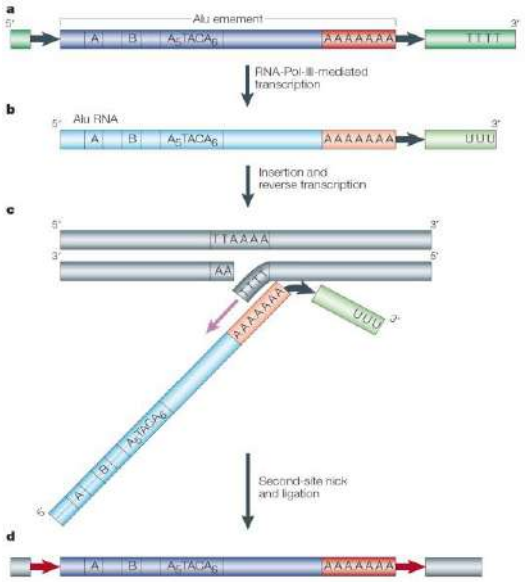


Figure 2. 9. Alu repeats and human genomic diversity.

The structure of each Alu element is bi-partite, with the 3′ half containing an additional 31-bp insertion (not shown) relative to the 5′ half. They also contain a central A- rich region and are flanked by short intact direct repeats that are derived from the site of insertion (black arrows). The 5′ half of each sequence contains an RNA- polymerase - III promoter (A and B boxes). The 3′ terminus of the Alu element almost always consists of a run of As that is only occasionally interspersed with other bases (a). Alu elements increase in number by retrotransposition — a process that involves reverse transcription of an Alu-derived RNA polymerase III transcript. As the Alu element does not code for an RNA-polymerase-III termination signal, its transcript will therefore extend into the flanking unique sequence (b). The typical RNA-polymerase-III terminator signal is a run of four or more Ts on the sense strand, which results in three Us at the 3′ terminus of most transcripts. It has been proposed that the run of As at the 3′ end of the Alu might anneal directly at the site of integration in the genome for target-primed reverse transcription (purple arrow indicates reverse transcription) (c). It seems likely that the first nick at the site of insertion is often made by the L1 endonuclease at the TTAAAA consensus site. The mechanism for making the second- site nick on the other strand and integrating the other end of the Alu element remains unclear. A new set of direct repeats (red arrows) is created during the insertion of the new Alu element (d) (Batzer and Deininger 2002).

3. OBJECTIVE

MobiSeq RRL protocol is now established as a target enrichment method for the recovery of non- human DNA sequence data through the targeted sequencing of the flanking regions of transposable elements (TEs) (Rey-Iglesia et al. 2018). However, taking into account the fragmented nature of ancient human DNA and the low endogenous DNA content in ancient samples, it remains to be seen if this novel method can be successfully applied in those samples.

This study aims to develop a targeting enrichment method based on the same principles with MobiSeq RRL protocol for non-human DNA. MobiSeq target enrichment approach is a PCR targeted-enrichment method generating sequence data from ancient samples using simple laboratory techniques. This method benefits from the “copy-and-paste” mobilization characteristic of retrotransposons, which will ensure that a large fraction of the sequenced loci can be compared among samples. Two different primers were selected to capture human genomic regions of interest.

20 human DNA extracts were targeted by two different TE-target primers for Alu gene families that showed hallmarks of recent activity. Shotgun libraries were also built to compare the number of single-nucleotide polymorphisms (SNPs) that could be recovered from 1k genomes with our method. MobiSeq target enrichment method showed high SNPs recovery using Alu_v1 TE-target primer for 4 of the human samples. Although, the clonality of this method seems to be really high. Thus, optimization of this method can be performed on TE-target primer design, titration, sequencing and computational analysis to reach the same level of coverage in human degraded samples. Finally, extended downstream analysis for Alu_v2 TE-target primer is required to fully evaluate this method.

4.PLAN OF ACTION

- Designing two TE primers for same SINE TE, Alu genes combined with the sequence of the P7 adapter Meyer and Kircher 2010, in order to create a fusion primer (TE+P7), which allows it to target only the Alu proximal sites.
- Designing a modified version of the P5 adapter from Meyer and Kircher 2010, the IS1 oligo is kept as in Meyer and Kircher 2010. IS3 oligonucleotide modified sequence by adding a C3 spacer at the 3' end blocking polymerase extension and a conventional lack of a 5'- phosphate.
- DNA extraction from human petrous bones using Allentoft et al. 2015 extraction protocol.
- Illumina library (MobiSeq libraries) building based on a recently developed blunt-end single- tube protocol (Carøe et al. 2018) with some modifications as in Mak et al. 2017.
- Quantification of MobiSeq Libraries using Real- time PCR (Mx3005P™ QPCR System) and a mix of SYBR Green/ ROXY dyes.
- Perform TE-target enrichment PCR on MobiSeq libraries for fragments containing the TE of interest.
- Indexing and amplification of TE-enriched libraries for sequencing as described in Meyer and Kircher 2010.
- Illumina library (Shotgun Libraries) building was based on the original blunt-end single-tube protocol (Carøe et al. 2018).
- Quantification of shotgun libraries using Real- time PCR (Mx3005P™ QPCR System) and a mix of SYBR Green/ ROXY dyes.
- Indexing and amplification shotgun libraries for sequencing as described in Meyer and Kircher 2010.
- Quantification and qualification of indexed shotgun library PCR products using TapeStation-High Sensitivity 2200 DNA (Agilent).
- Sequencing of both indexed libraries (MobiSeq and Shotgun libraries) at the Danish National High-throughput Sequencing Centre, Copenhagen, Denmark, on an Illumina HiSeq Instrument for 81 cycles in single-end read mode.
- Evaluation and comparison of both sequences (MobiSeq and shotgun sequences) to obtain if the recovery of the SNPs and endogenous DNA are successful using this target enrichment protocol.

5. MATERIALS AND METHODS

5. 1. Sample information

DNA was extracted from 20 human petrous bones, see Table S1- Appendix 1 for a detailed description of the specimens included in the study. The individuals used in the study were found in central Copenhagen and they are dated between 16th- 18th century. The samples were ideal for this study because they weren't extremely old and they were also well preserved.

DNA extractions were performed using a silica-based extraction protocol (Allentoft et al. 2015). DNA elution was performed twice in 30 µl EB buffer and with 10 minutes of incubation time at 37 °C prior to elution, in order to increase DNA yield. Extractions were quantified using a Qubit dsDNA High Sensitivity (HS) assay (Life technologies). Additionally, we used modern human DNA as positive control which was also quantified using a Qubit dsDNA High Sensitivity (HS) assay (Life technologies) and TapeStation High Sensitivity DNA 2200 (Agilent).

5. 2. TE-target primer design

In order to enrich our libraries for specific TE-target elements, we designed two TE primers for repeat families that showed hallmarks of recent activity. Highly conserved regions were determined from the alignments and selected as potential primer sites. This resulted in two primers targeting the SINE TE AluY gene subfamily (Alu_v1: 5′-ATTACAGGCGTGAGCCACCGCGCC-3′ and Alu_v2: 5′-TGAGCCACCGCGCCCGGC-3′). The TE-target oligonucleotides were then combined with the sequence of the P7 adapter (GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT) (Meyer and Kircher 2010), in order to create a fusion primer (TE+P7) that would enrich for TE sites, at the same time as adding the P7 sequencing adapter compatible with binding to Illumina flow cells (see Table S2, Appendix S1 for an overview of the oligonucleotides).

5. 3. Modified P5 adapter

A modified version of the P5 adapter from Meyer and Kircher 2010 was designed for this protocol. In this modified P5 (mp5), the IS1 oligo is kept as in Meyer and Kircher 2010. However, IS3 oligonucleotide presents a modified sequence by adding a C3 spacer at the 3' end blocking polymerase extension. This, together with the conventional lack of a 5'-phosphate, allows us to run a PCR reaction using a universal primer for the adapter sequence (IS4, see Table S2, Appendix S1) and a TE-target primer enriching for a specific subset of TEs. Hybridization of IS1 and the modified IS3 to generate mp5 was performed as in Meyer and Kircher 2010.

5. 4. Library preparation

5. 4. 1. MobiSeq libraries

5. 4. 1. 1. End-repair and adapter ligation

A schematic overview of the method is represented in Figure S1- Appendix 1 and the detailed protocol is available as Appendix S2.

Illumina targeted library (MobiSeq libraries) building was based on a recently developed blunt-end single-tube protocol (Carøe et al. 2018) with some modifications as in Mak et al. 2017 (full protocol available as Appendix S2). In particular, the ligated adapters differed from the ones in the original protocol by excluding the use of a P7 adapter (only using the mp5 adapter) and excluding the adapter fill-in reaction. Following library preparation, the reactions were purified using a magnetic beads purification protocol, Sera- Mag Speedbeads (Thermo Scientific), at 1.8x. Purified libraries were eluted in 30 µl of EB.

The positive control was fragmented prior to library build as the average DNA fragment size was ca. 350- 400 bp. Starting material for Bioruptor was 347 ng depending on the positive control concentration. The length profile of the fragmented material was obtained using TapeStation High Sensitivity (TS- HS) DNA 2200 (Agilent) (results in Table S3-1, Appendix S3).

5. 4. 1. 2. Quantification and quality control

Libraries were quantified using Real- time PCR (Mx3005P™ QPCR System) and a mix of SYBR Green/ROXY dyes (full protocol available as Appendix S2). QPCR

reactions were performed in 20 µl containing: 1x KAPA HiFi HotStart Uracil+ ReadyMix, 0.6 µM of IS4 primer, 0.6 µM of Alu_v1 or Alu_v2 primer 1% BSA (Sigma-Aldrich), 0.8 µl of mix of SYBR Green/ ROXY dyes and 1 µl of the purified libraries. Cycling parameters were denaturation at 95°C for 3 minutes, followed by 35 cycles of denaturation at 98°C for 20 seconds, annealing at 65 °C (depending on the TE-target primer) 15 seconds, and extension at 72°C for 30 seconds) and an additional extension at 72°C for 3 minutes. The TE PCR cycles of amplification were measured according to Cp/ Ct values of the qPCR results (Table S3a- Appendix S1).

5. 4. 1. 3. TE-enrichment PCR

Libraries were enriched for fragments containing the TE of interest by using a TE target enrichment PCR. Primers for this PCR were forward primer IS4 (Meyer and Kircher 2010) and the fusion reverse primer described in prior sections (Table S2 for oligonucleotide sequences, Appendix S1). PCR reactions were performed in 20 µl containing: 1x KAPA HiFi HotStart Uracil+ ReadyMix, 0.3 µM of each primer, 2% BSA (Sigma-Aldrich), and 5 µl of the purified libraries. Cycling parameters were denaturation at 95°C for 3 minutes, followed by 22-30 cycles (according to qPCR Ct/Cp values) of denaturation at 98°C for 20 seconds, annealing at 65°C (depending on the TE-target primer) 15 seconds, and extension at 72°C for 30 seconds) and an additional extension at 72°C for 3 minutes. TE- enriched libraries were purified using magnetic bead solution at 1.8x magnetic bead solution. Purified DNA was eluted in 30 µl of EB.

5. 4. 1. 4. Indexing PCR

TE-enriched libraries were indexed and amplified for sequencing as described in Meyer and Kircher 2010. PCR reactions were performed in 20 µl containing: 1x KAPA HiFi HotStart Uracil+ ReadyMix, 0.3 µM of each primer, 2% BSA (Sigma-Aldrich), and 5 µl of the purified libraries. Cycling parameters were denaturation at 95°C for 3 minutes, followed by 5-10 cycles of denaturation at 98°C for 20 seconds, annealing at 65°C for 15 seconds, and extension at 72°C for 30 seconds) and an additional extension at 72°C for 3 minutes. Indexed libraries were purified using magnetic bead solution at 1.8x of magnetic bead solution. Purified DNA was eluted in 30 µl of EB. Concentration was measured TapeStation-High Sensitivity DNA 2200 (Agilent).

5. 4. 2. Shotgun libraries

5. 4. 2. 1. Adapter ligation, end repair and adapter fill-in

Illumina shotgun library (Shotgun Libraries) building was based on the original blunt-end single-tube protocol (Carøe et al. 2018). Following library preparation, the reactions were purified using a magnetic beads purification protocol, Sera-Mag Speedbeads (Thermo Scientific), at 1.8x of magnetic bead solution. Purified libraries were eluted in 30 µl of EB.

5. 4. 2. 2. Quantification and quality control

Libraries were quantified using Real- time PCR (Mx3005P™ QPCR System) and a mix of SYBR Green/ROXY dyes (full protocol available as Appendix S1, Supplementary Material). qPCR reactions were performed in 20 µl containing: 2 µl AmpliTaq Gold Buffer, 2 µl AmpliTaq Gold $MgCl_2$, 0.4 µM of IS4 primer, 0.4 µM of reverse primer, 2% BSA (Sigma- Aldrich), 0.8 µl of mix of SYBR Green/ ROXY dyes, 0.16 µl of dNTPs (25 mM), 0.16 µl of AmpliTaq Gold Polymerase and 1 µl of the purified libraries. Cycling parameters were denaturation at 94°C for 4 minutes, followed by 35 cycles of denaturation at 94°C for 30 seconds, annealing at 58°C for 30 seconds, and extension at 72°C for 30 seconds. The PCR cycles of amplification were measured according to Cp/ Ct values of the qPCR results (Table S3b, Appendix S1).

5. 4. 2. 3. Indexing PCR

Shotgun libraries were indexed and amplified for sequencing as described in Meyer and Kircher 2010. PCR reactions were performed in 50 µl containing: 1x Taq Gold Buffer, 2.5 mM of $MgCl_2$, 0.4 µM of each primer, 0.1 mM dNTPs, 2% BSA (Sigma-Aldrich), 0.02 U/µl AmpliTaq Gold DNA polymerase and 10 µl of the purified libraries. Cycling parameters were denaturation at 94°C for 4 minutes, followed by 19-32 cycles (according to qPCR Cp/Ct values) of denaturation at 94°C for 30 seconds, annealing at 58°C for 30 seconds, and extension at 72°C for 30 seconds. Indexed libraries were purified using magnetic bead solution at 1.8x of magnetic bead solution. Purified DNA was eluted in 30 µl of EB. Concentration was evaluated with TapeStation-High Sensitivity DNA 2200 (Agilent).

5. 4. 3. Sequencing

Indexed libraries were pooled, giving 30% of the lane to MobiSeq and 70% to shotgun, and sequenced at the Danish National High-throughput Sequencing Centre, Copenhagen, Denmark, on an Illumina HiSeq Instrument for 81 cycles in single read mode. Human reference genome is wide, and probably has a large chunk of the Alu sites in it already. Thus, it is feasible with quite high confidence ascertain the location of the read and its distance from an Alu site. The sequencing architecture is illustrated in Figure S2, Appendix S1**.**

5. 4. 4. Data processing

5. 4. 4. 1. MobiSeq libraries data analysis

Single-end reads were trimmed of adapter Illumina sequences and reads shorter than 25 bp were discarded and quality filtered using AdapterRemoval v2.0 (Schubert, Lindgreen, and Orlando 2016). Compared to the original pipeline (Rey-Iglesia, Gopalakrishnan, and Carøe 2018), the presence of the TE-target primer sequence was verified in the 3' end of the single read. Then, we checked primer presence using cutadapt v2.3 (Martin 2011). The sequenced bases reads were 80 bp SR, so the reverse complement of the primer will be now in the 3' of the R1.

As no PE read sequencing mode was performed it is unsure if all the R1 contain the primer. Therefore, all the generated reads after adapter removal were separately mapped against the human reference genome. Mapping was performed with BWA-v0.7.15 aln (Li and Durbin 2010). PCR duplicates and reads mapping to multiple genomic locations were marked using SAMtools-v1.6 (Li et al. 2009). Reads mappings were employed for the downstream analysis.

The authenticity of the reads was tested characterizing the presence of post-mortem damage (Willerslev and Cooper 2005). Those patterns in ancient DNA sequences were tracked and quantified damage using the mapDamage tool (Jónsson et al. 2013).

SNPs discovery across 1k genome data set for MobiSeq libraries was performed using samtools to retrieve the variable sites and then using bedtools (Quinlan and Hall 2010) to compare those variable sites to SNPs presented in MobiSeq libraries.

5. 4. 4. 2. Shotgun libraries data analysis

Basic sequencing statistics, such as read numbers and clonality were quantified within the PALEOMIX pipeline (Schubert et al. 2014). First, single-end reads were trimmed of adapter Illumina sequences and reads shorter than 25 bp were discarded and quality filtered using AdapterRemoval v2.0 (Schubert, Lindgreen, and Orlando 2016). Then all generated reads were separately mapped against the human reference genome. Mapping was performed with BWA-v0.7.15 aln (Li and Durbin 2010). PCR duplicates and reads mapping to multiple genomic locations were marked using SAMtools-v1.6 (Li et al. 2009). We tested the authenticity of the mapped reads using the mapDamage pipeline (Jónsson et al. 2013). Finally, SNPs discovery across 1k genome data set for shotgun libraries was performed using samtools to retrieve the variable sites and then using bedtools (Quinlan and Hall 2010) to compare those variable sites to SNPs presented in shotgun libraries.

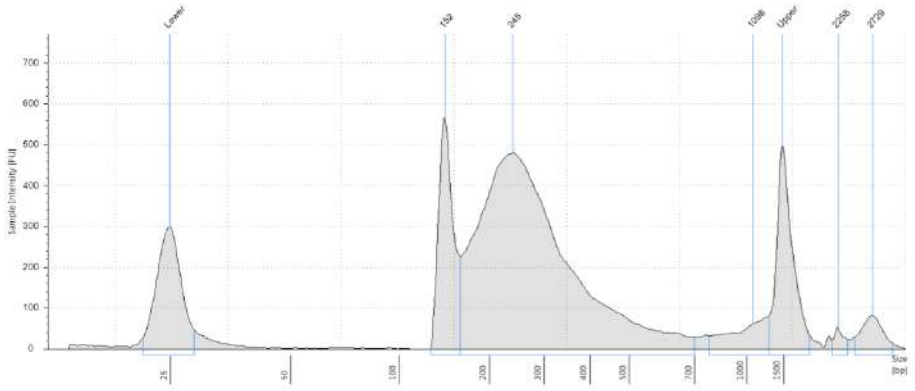
6. RESULTS

6. 1. TapeStation results for MobiSeq and shotgun libraries

Figure 6. 1. presents the results of quality and quantity control for MobiSeq libraries for P63, P73 individuals and positive controls processed with both Alu-TE target primers. We obtain long fragments of DNA in P63 and in the positive control MobiSeq libraries targeted with both Alu TE-target primers that are chimeric products of the PCR amplification. In addition, in some samples a shorter DNA fragment peak appeared at approximately 80bp, those fragments are primer duplicates. Due to the high molecular weight DNA fragments, accurate measurement of library concentration can not be performed using TapeStation High Sensitivity DNA 2200 (Agilent). Non accurate concentration estimation creates titration issues in sequencing by generating massive number of reads for the same individual.

However, P73 showed the expected library profile with DNA fragments approximately at 200bp, given the expected ancient human DNA fragment is ~80bp, TE-target primer 18-24bp, Illumina adapters ~40bp each and dual indexes ~6bp each. Also, the concentration measurement was accurate.

B2: P63 alu-v2 9/5 indexed MobiSeq Libs



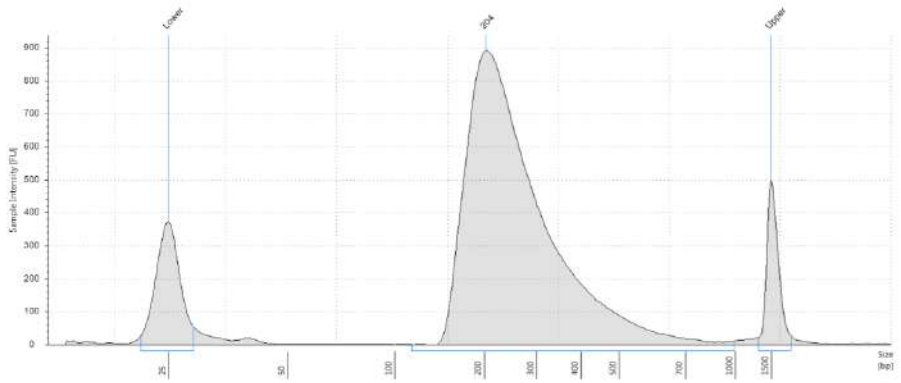
Sample Table

Well	Conc. [pg/µl]	Alert
B2	1750	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	232	-	14300	-		Lower Marker
152	264	-	2670	15.03		
245	1350	-	8440	76.70		
1098	78.6	-	110	4.48		
1500	250	250	256	-		Upper Marker
2258	14.2	-	9.69	0.81		

D1: P73 aluv1 index PCR product



Sample Table

Well	Conc. [pg/µl]	Alert
D1	2860	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	385	-	23700	-		Lower Marker
204	2860	-	21600	100.00		
1500	250	250	256	-		Upper Marker

Figure 6. 1. TapeStation profiles for P63, P73 individuals using both primers.

6. 2. Number of sequenced reads and clonality for MobiSeq libraries

All samples for MobiSeq protocol were processed at the same time and date, apart from P63 and P73 using both primers. The results of this study present a high variability. Our sequencing yielded a total of 290.664.569 raw reads for both libraries (MobiSeq and shotgun). The range of total reads per sample is between 835.827 and 23.557.045 reads.

For MobiSeq libraries we generated 224.838.269 filtered reads (after adapter removal). 25% of the reads were discarded after trimming Illumina adapters and quality filtering, and an average of 79% of the remaining reads with the TE-target primer were mapped to the human genome (detailed data in Table S4, Appendix S1). The samples used for each one of the primers were different, but randomly selected for each TE-target primer.

6. 3. Comparison of two different Alu TE-target primers

6. 3. 1. Comparison of the Alu TE-target primers using different ancient individuals

Comparing the targeted performance and the clonality of both Alu TE-target primers, Alu_v1 generated less total reads (Alu_v1: 4.070.882 and Alu_v2: 12.983.389 number of total reads), and the filtered reads after removing the adapters are consequently decreased (Alu_v1: 3.941.522 and Alu_v2: 12.098.769 number of filtered reads) compared to Alu_v2 TE-target primer (Figure 6. 2). (detailed information in Table S5- Appendix S1).

The averages of mapped and unique reads (TE-target primer present and absent) for Alu_v1 TE-target primer are 8.461.510 and 4.059.067, accordingly. Additionally, the average of mapped reads for Alu_v2 TE-target primer is 9.710.537 and the average of unique reads is for Alu_v2 TE-target primer (TE-target primer present and absent). The endogenous content (%) and the clonality (%) per Alu TE-target primers are presented in Figure 6. 3. Alu_v1 TE-target primer gives higher endogenous DNA (76%) slightly less PCR duplicates (75%) than Alu_v2 (61% and 76%, accordingly).

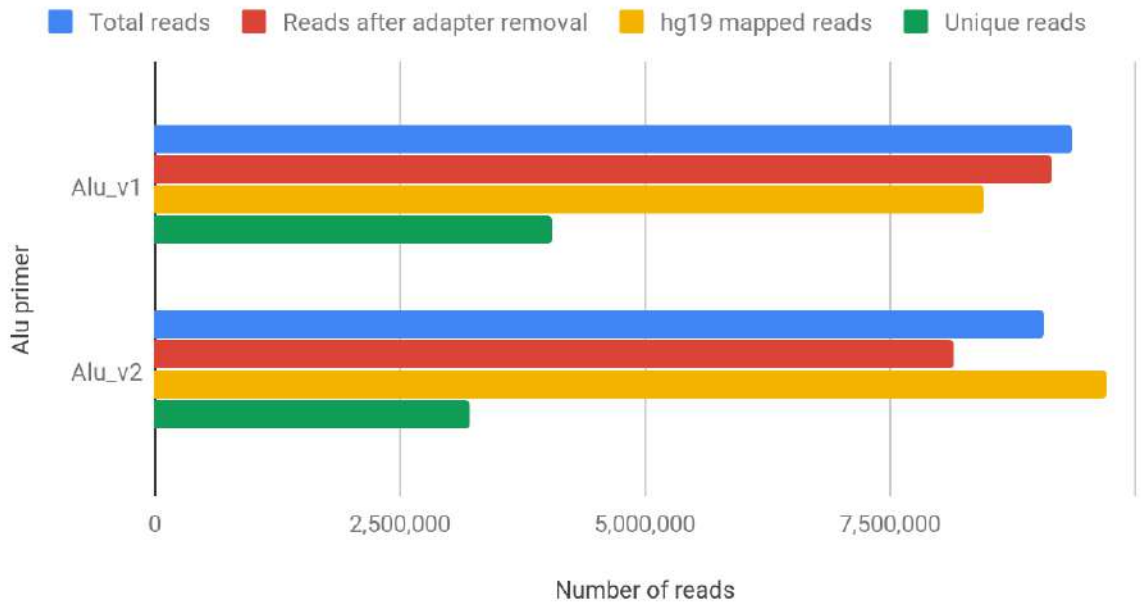


Figure 6. 2. Average sequencing and mapping statistics for Alu TE-target primers.

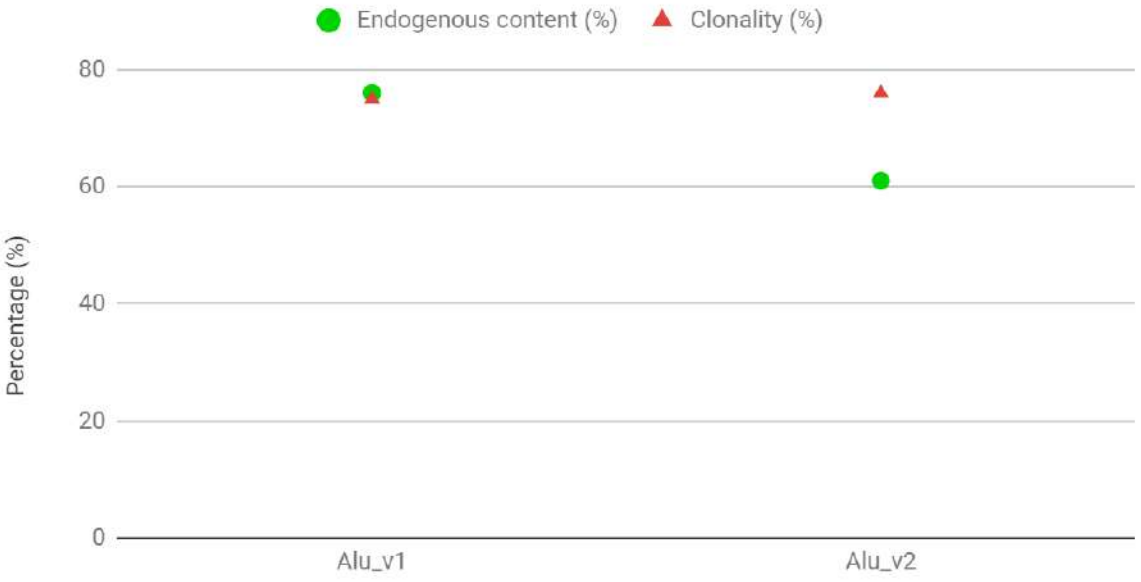


Figure 6. 3. Endogenous content (%) and clonality (%) for Alu TE-target primers.

6. 3. 2. Comparison of the Alu TE-target primers using the same ancient individual

In general terms, the results presented that sample P63 has lower quality than P73. However, an assumption can be made using the sample P73. Alu_v1 in sample P73 generated more sequenced data for the region of interest (Figure 6. 4).

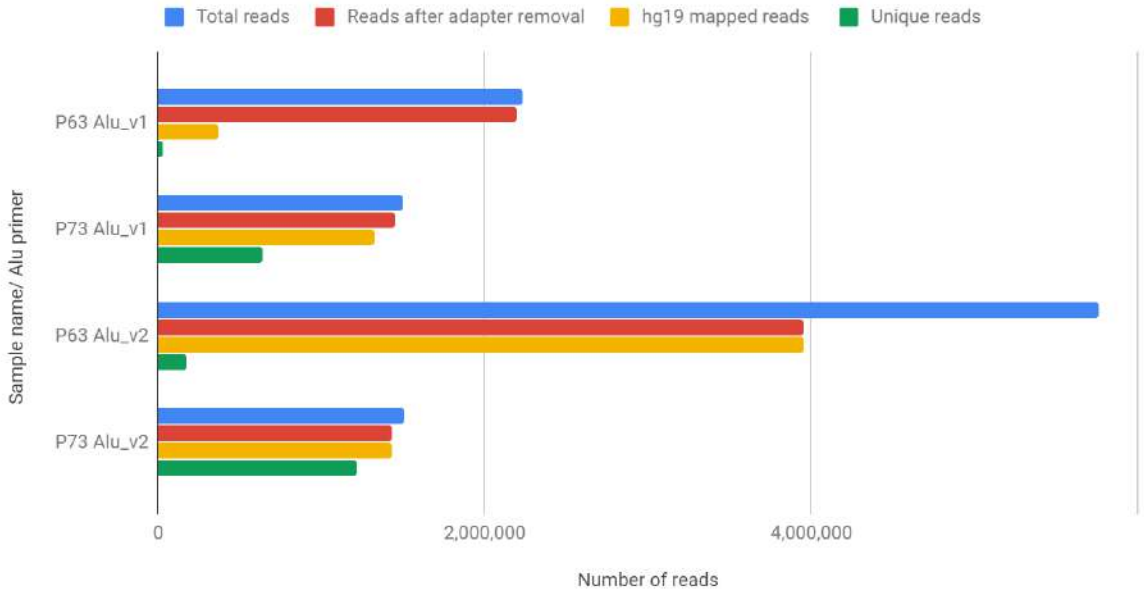


Figure 6. 4. Average sequencing and mapping statistics for two different ancient samples (mobiSeq libraries) using both Alu TE-target primers.

As expected considering the quality of the starting material, endogenous DNA recovery is higher in P73 using both primers (Alu_v1: 89%, Alu_v2: 80%) than in P63 (Alu_v1: 85%, Alu_v2: 3%). The clonality seems to be correlated with the quality of the DNA, too. Usually more PCR cycles generate more duplicates. In addition, most of the reads with both the Alu TE-target primers mapped to the human genome (P63 Alu_v1: 98%, Alu_v2: 86% and P73 Alu_v1: 52%, Alu_v2: 63%). Alu_v1 TE-primer generated more mapped reads than Alu_v2 in both samples. Comparing those two individuals Alu_v2 TE-target primer appears to generate higher clonality rates than Alu_v1 TE-target primer. However, in P73 processed with Alu_v1 the endogenous DNA recovery was slightly poorer than the same sample with Alu_v2 (Figure 6. 5).

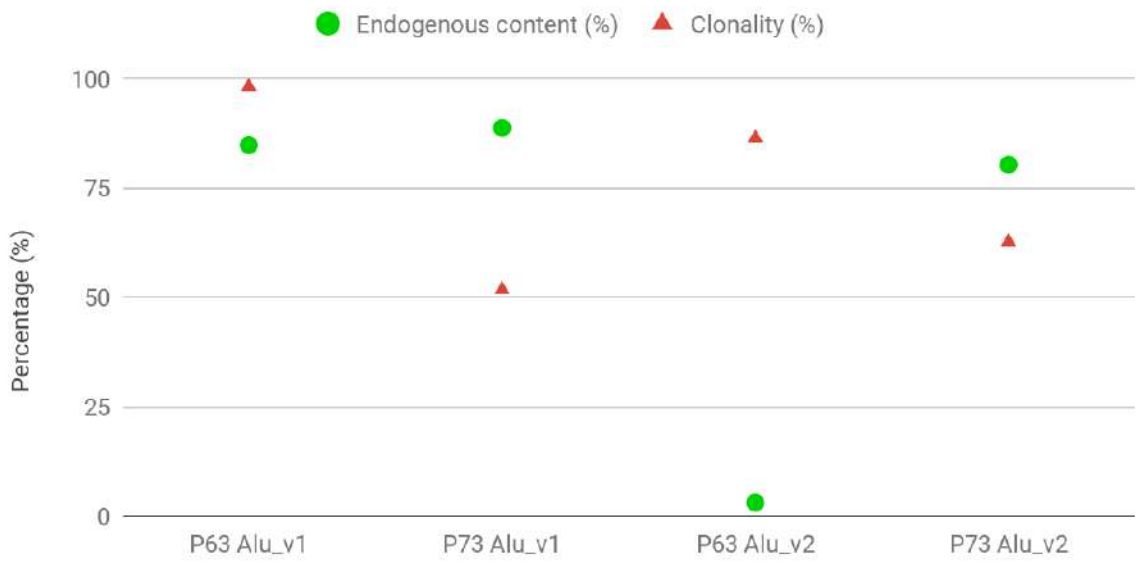


Figure 6. 5. Endogenous content (%) and clonality (%) for two different ancient samples (mobiSeq libraries) using both Alu TE-target primers.

6. 3. 3. Comparison of the Alu TE-target primers using modern DNA

The number of total reads for modern samples is higher compared to the ancient samples. The concentration of the samples was not accurately quantified due to the long DNA fragments presented in modern samples. This seems to have caused an unbalanced titration on the sequence pool, generating high amount of sequenced data that took over parts of the ancient sequence data. Nevertheless, MobiSeq performance in modern human samples with Alu_v2 TE-target primer generated slightly more sequenced data than Alu_v1. But, higher clonality were obtained using Alu_v2 TE-target primer.

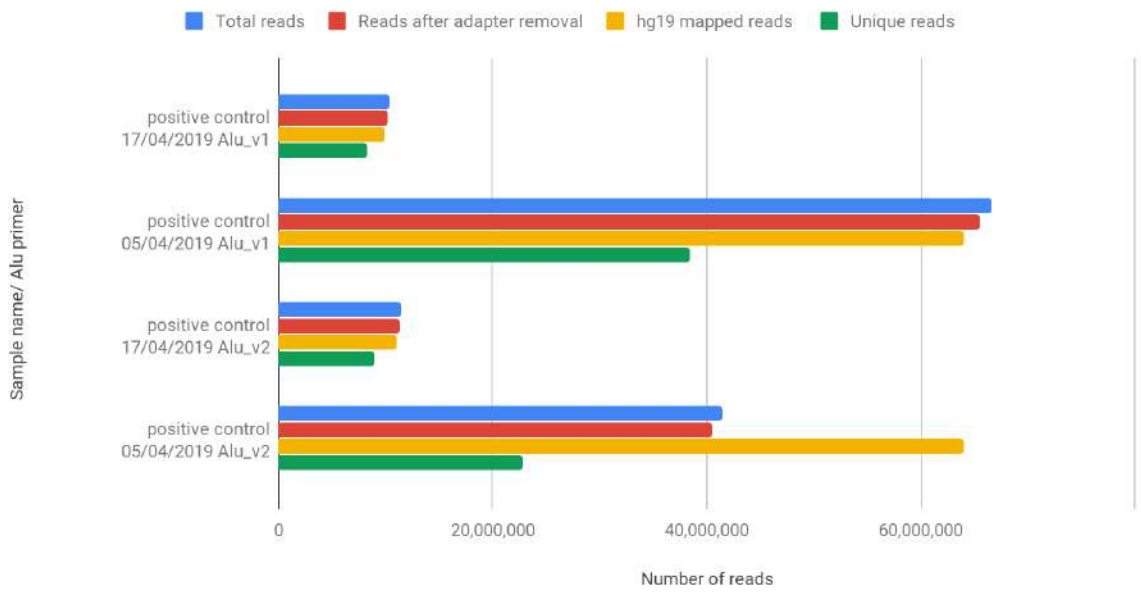


Figure 6. 6. Average sequencing and mapping statistics for modern human DNA mobiSeq libraries using both Alu TE-target primers.

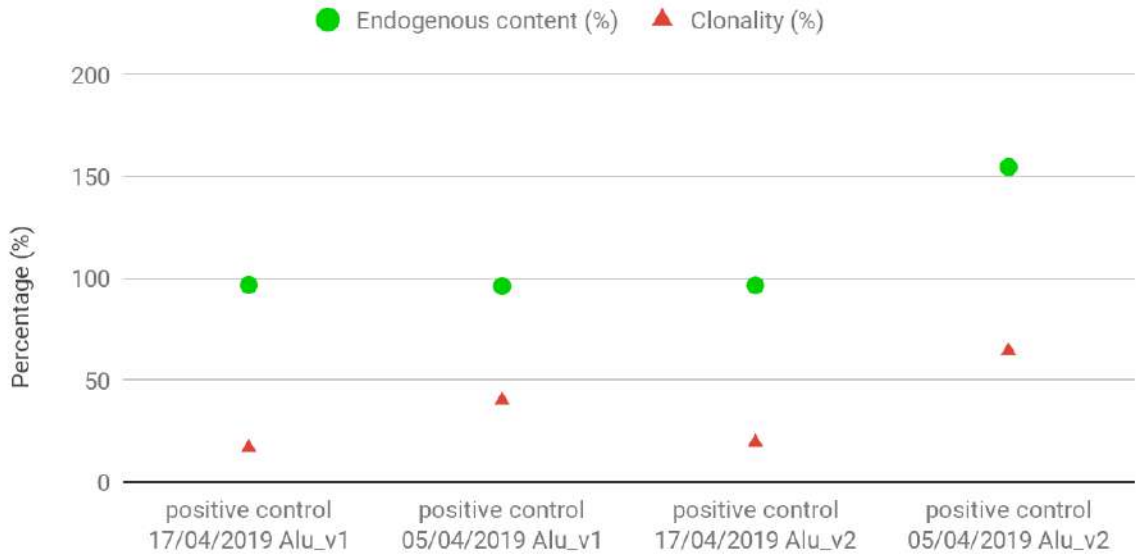


Figure 6. 7. Endogenous content (%) and clonality (%) for modern human DNA mobiSeq libraries using both Alu TE-target primers.

6. 4. Number of sequenced reads and clonality for shotgun libraries

All samples for shotgun libraries were processed at the same time and date. For shotgun libraries we generated 51.134.032 total reads and 49.763.642 filtered reads (after adapter removal). Two of the samples yielded extremely low coverage of DNA

sequenced data due to a mix of indexing primers. In Figure 6. 9 we present the total number of reads, the filtered reads (after adapter removal), the reads that mapped to the human reference genome and finally the unique DNA sequences with the valid TE-target primer for shotgun libraries (detailed information in Table S6- Appendix S1).

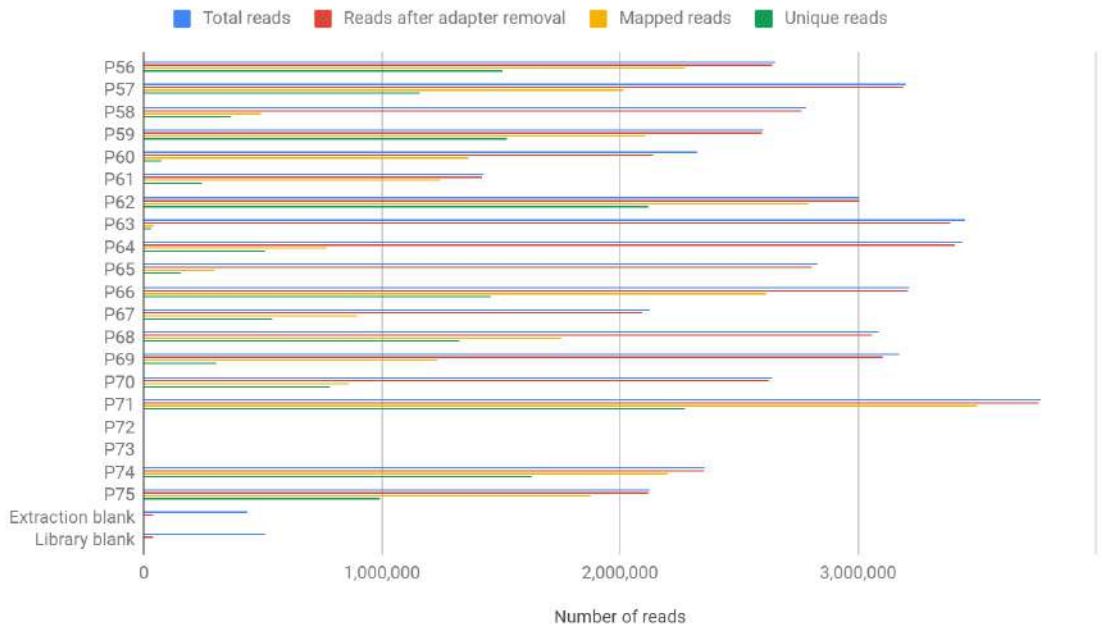


Figure 6. 9. Average sequencing and mapping statistics for shotgun libraries.

Shotgun libraries present high clonality rates but also high endogenous content. Excluding extraction and library building blank there is an average of 61% endogenous content, 38% of library efficiency and 63% of clonality (Figure 6. 10). (additional information in Figure S4- Appendix S1).

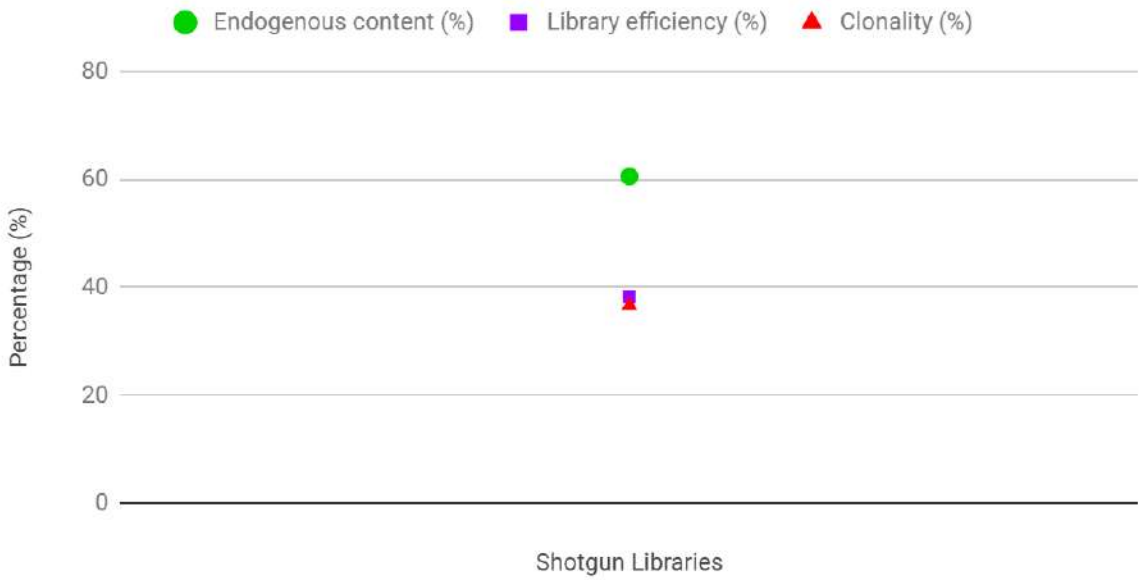


Figure 6. 10. Endogenous content (%), library efficiency (%) and clonality (%) for shotgun libraries.

6. 5. Authenticity of MobiSeq and Shotgun libraries.

6. 5. 1. MobiSeq Libraries

The authenticity of mapped reads was evaluated by tracking post- mortem damage (deamination, oxidation, depurination) and length distribution per individual. On this section we present the damage profile of the samples and the length distribution across 5 individuals. Figure 6. 11 shows the damage patterns at the 5' and 3' end for each one of the DNA samples. The length distribution per sample appears in Figure 6. 12 (fragment misincorporation and length plots for all of the samples are in Appendix S4).

A weird profile is obtained in both graphs on the 3' end of each fragment including the positive controls, due to the existence of the TE-target primer at the 3' end of each DNA fragment. This is also identified by the length plots of all the samples which showed that the sequenced DNA fragments are approximately 35 bases.

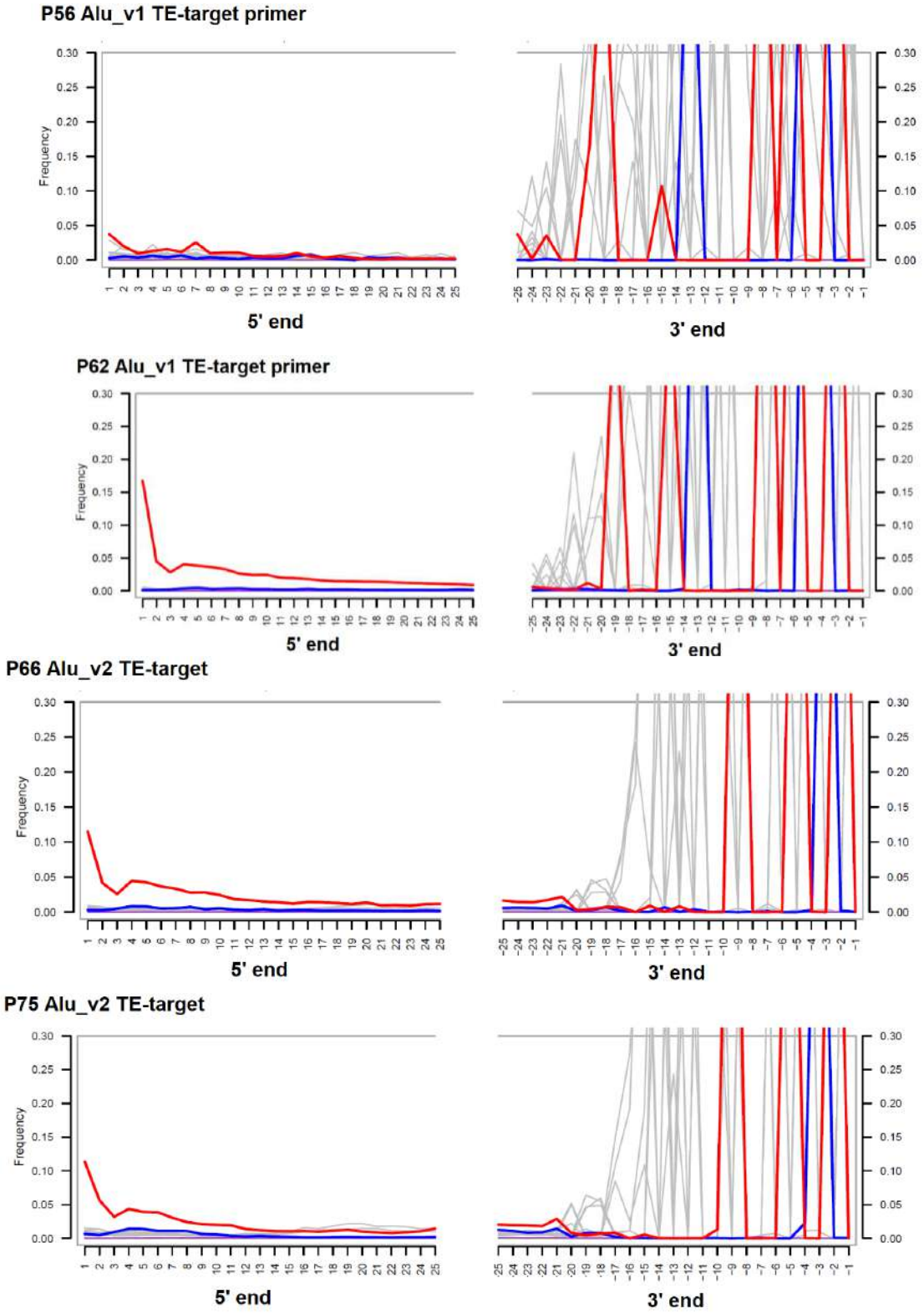


Figure 6. 11. Fragment misincorporation plots for 6 individuals using both Alu TE-target primers. The C to T transition of the DNA strand is represented by the red colour. The G to A transition is shown by the blue colour.

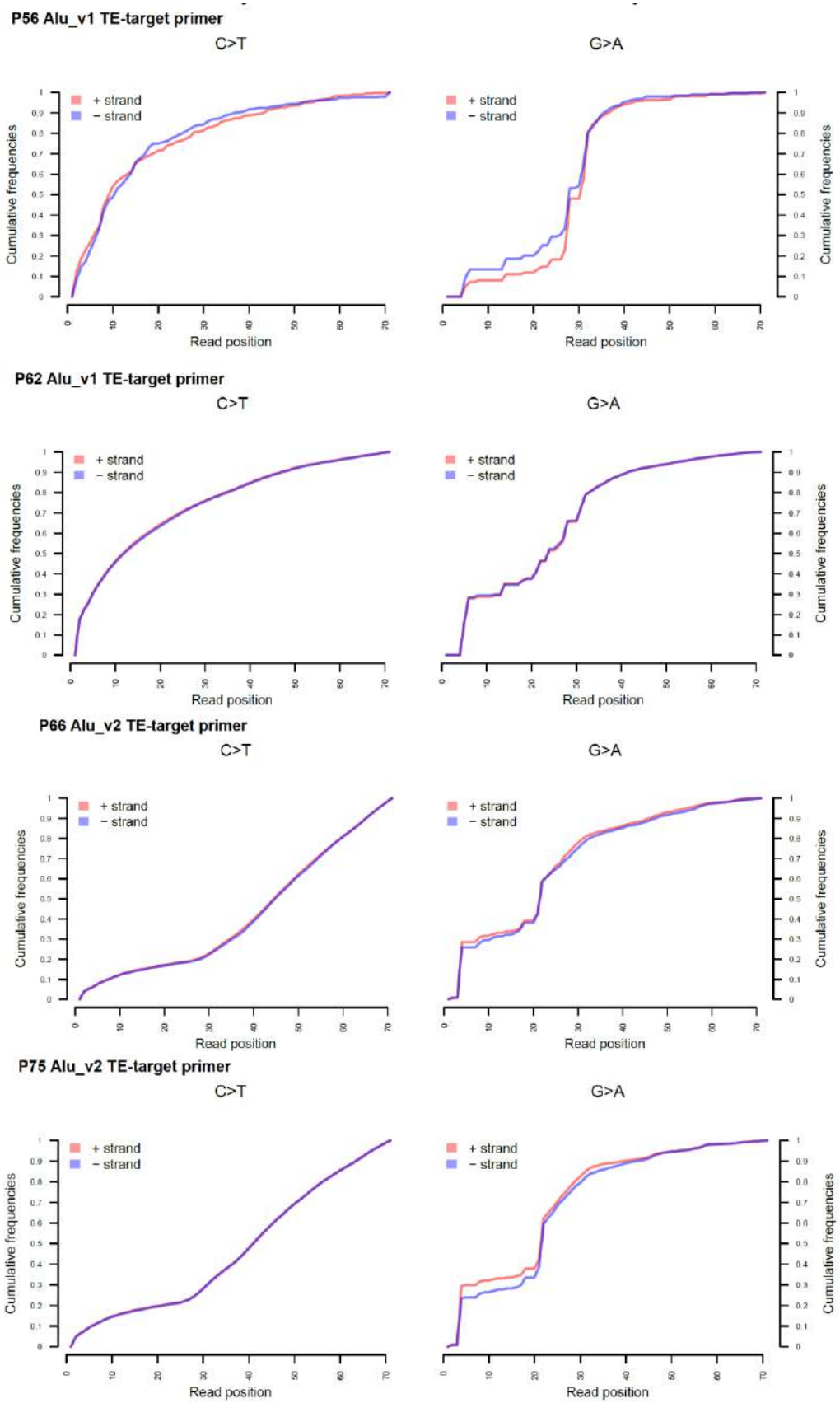


Figure 6. 12. Length plots for 6 individuals using both Alu TE-target primers.

In Figure 6. 13 and 6. 14 we present the fragment misincorporation plots and the length distribution for P63 and P73 individuals using both Alu TE-target primers accordingly. P63 shows no damage and high background noise in 5’ end, probably because of the really low recovery of DNA. However, the 3’ end has the same pattern for all of the samples. In addition, the expected end damage pattern appears in the sample P73 using both TE-target primers.

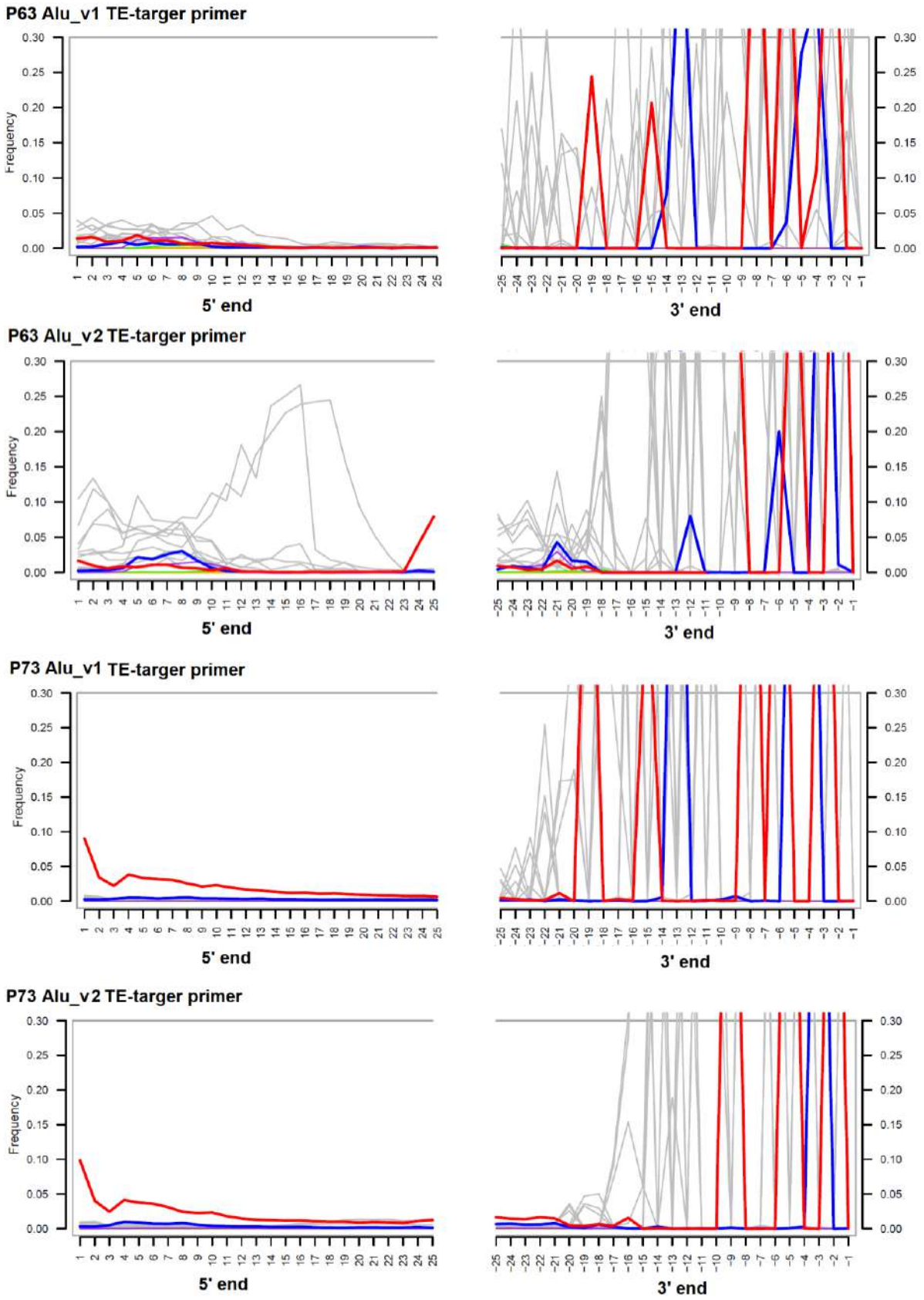


Figure 6. 13. Fragment misincorporation plots for samples P73 using both Alu TE-target primers.

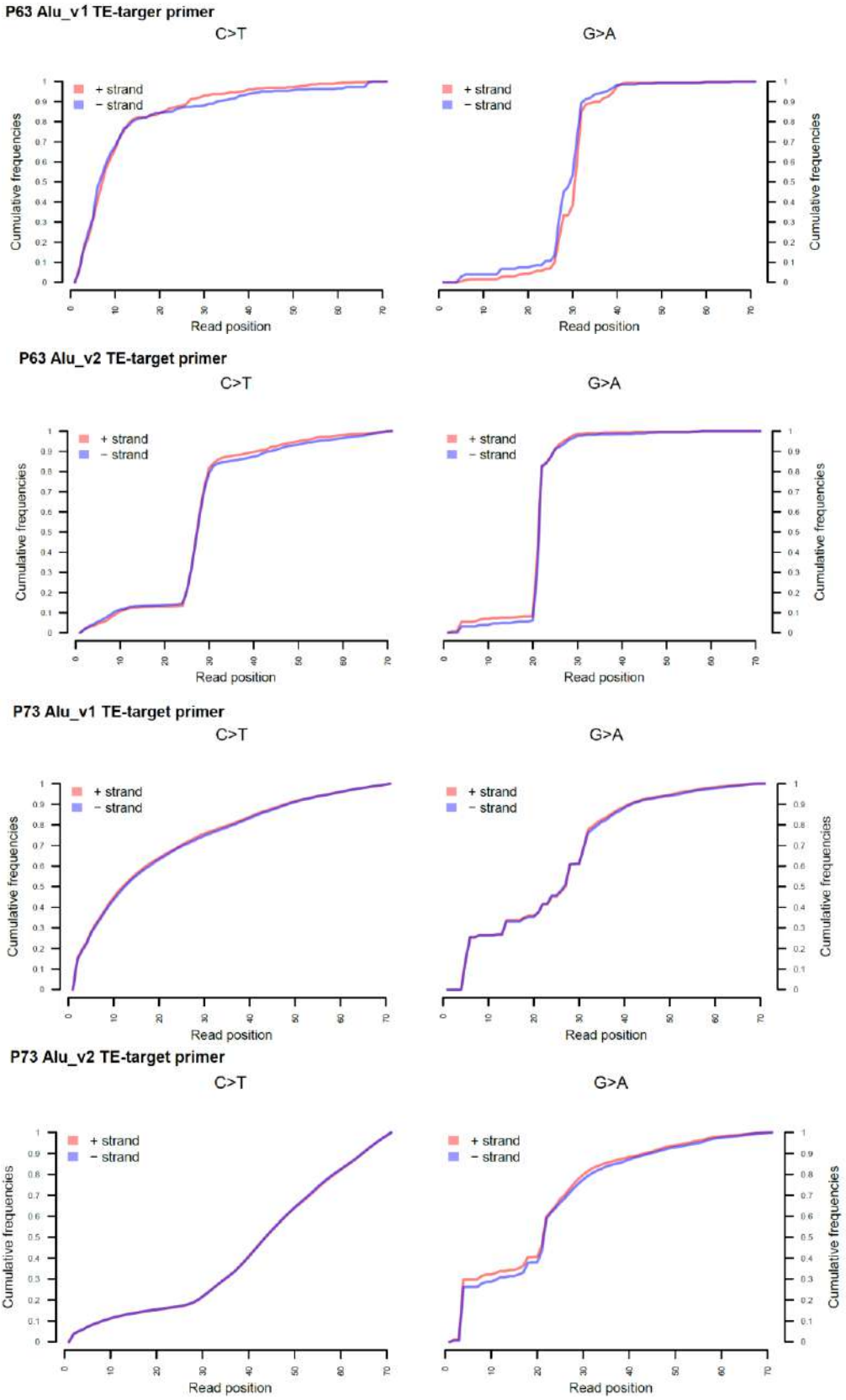


Figure 6. 14. Length plots for samples P63 and P73 using both Alu TE-target primers.

6. 5. 2. Shotgun Libraries

Figure 6. 15 shows the fragment misincorporation plots and the length distribution for one shotgun library of sample P68. All shotgun libraries obtain the same misincorporation and length distribution profile apart from individuals P72 and P73 (fragment misincorporation and length plots for all of the samples in Appendix S4). Those two samples generated very few sequenced DNA data that are not enough for downstream analysis.

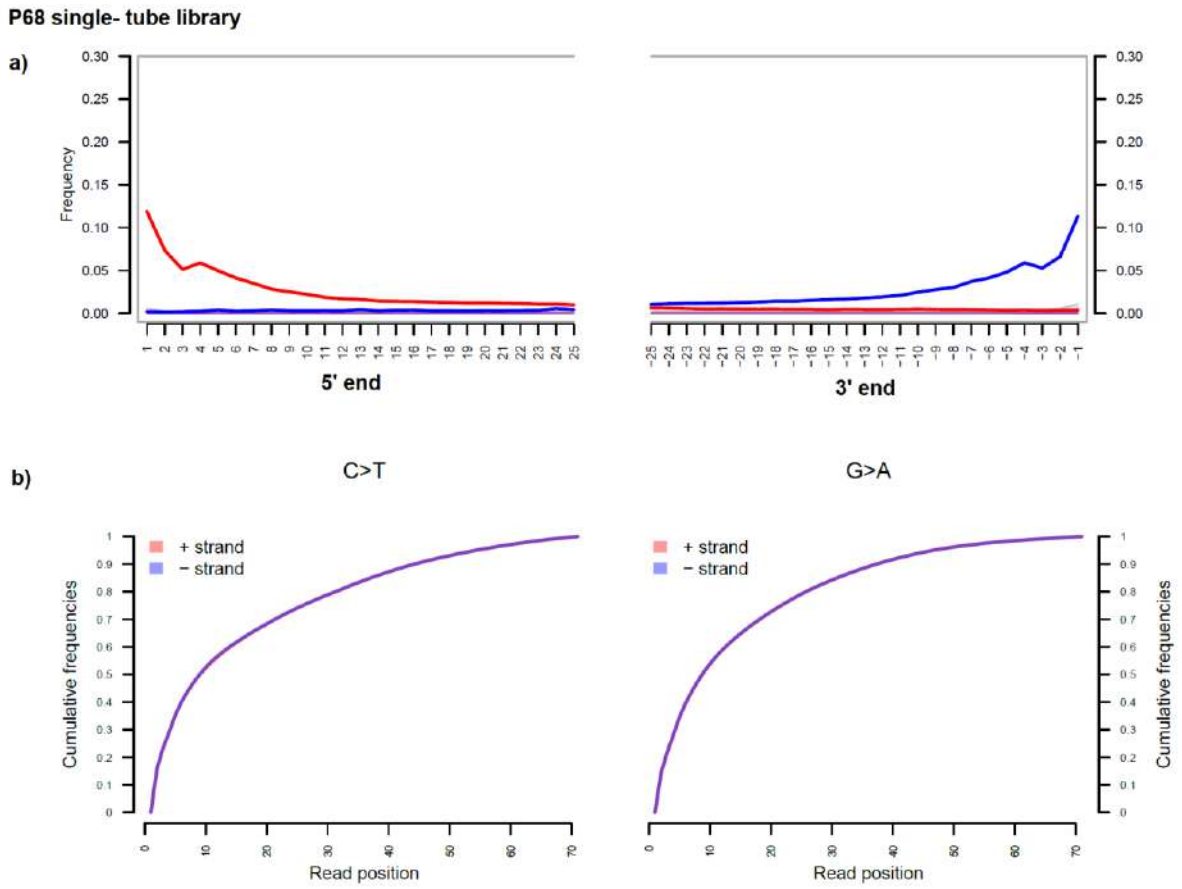


Figure 6. 15. a) Fragment misincorporation plot and b) length plot for sample P68 shotgun library.

6. 6. SNPs in 1k genome for Mobiseq and shotgun libraries.

For MobiSeq libraries a different computational approach than the one described in Rey-Iglesia et al. 2018 generated a more accurate set of results including SNPs number in 1k genomes. We mapped all the reads including those with no TE-target primer to the human reference genome. The analysis that search for SNPs only in reads with the TE primer was not efficient in this study. Due to the sequencing reads mode most of the reads containing the TE-target primer were discarded because of the sequencing read mode.

Figure 6. 16. shows the number of SNPs recovered from MobiSeq libraries for each one of the individuals using Alu_v1 TE-target primer, excluding positive controls. The average of SNPs captured with MobiSeq target method using Alu_v1 TE-target primer is 406.704 and the range is between 15.428 to 1.723.970 when the average number for shotgun libraries is 742.233 and the range is from 350 up to 3.300.570. However, there is a bias in shotgun libraries because almost no data were generated for sample P73 (detailed information in Table S7- Appendix S1).

The averages of endogenous content (%) and clonality (%) for Alu_v1 TE-target primer are 71% and 75% accordingly including the positive controls (Figure 6. 17).

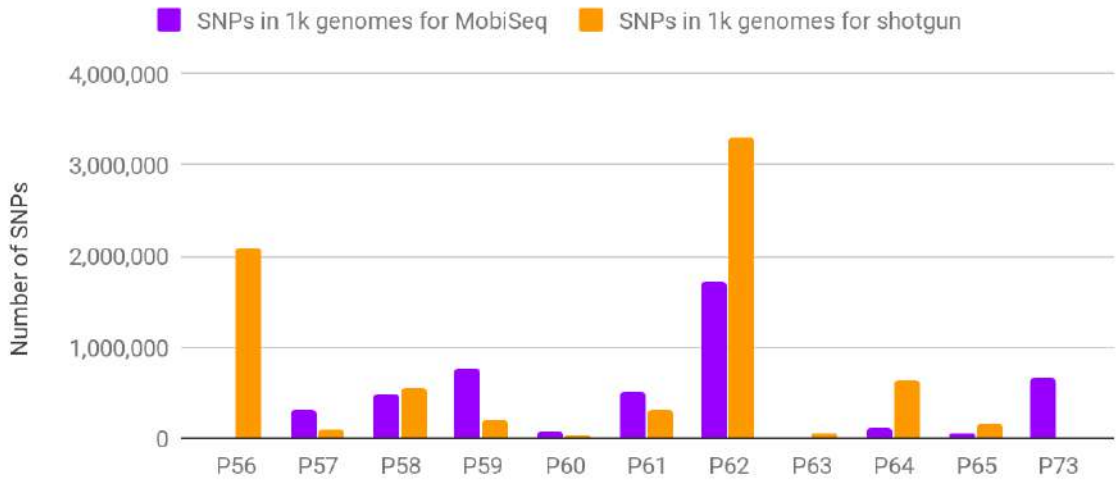


Figure 6. 16. SNPs recovery in 1k genomes for MobiSeq (Alu_v1) and shotgun libraries.

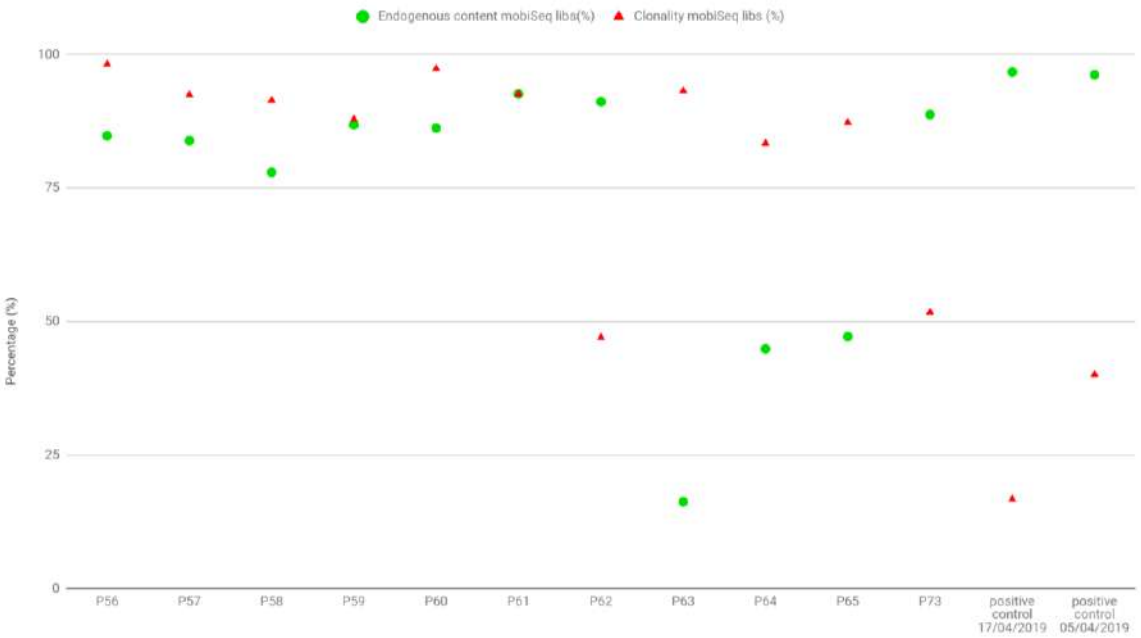


Figure 6. 17. Endogenous content (%) and clonality (%) for MobiSeq Alu_v1 TE-target primer libraries.

Figure 6. 18. Presents the number of SNPs recovered from MobiSeq libraries for each one of the individuals using Alu_v2 TE-target primer, excluding positive controls. The average of SNPs captured with MobiSeq target method using Alu_v2 TE-target primer is 332.061 and the range is between 3.094 to 1.187.156 when the average number for shotgun libraries is 1.191.164 and the range is from 350 up to 3,591,804. However, there is a bias in shotgun libraries because almost no data were generated for samples P72 and P73 (detailed information in Table S7-Appendix S1).

The averages of endogenous content (%) and clonality (%) for Alu_v1 TE-target primer are 61% and 76% accordingly including the positive controls (Figure 6. 19).

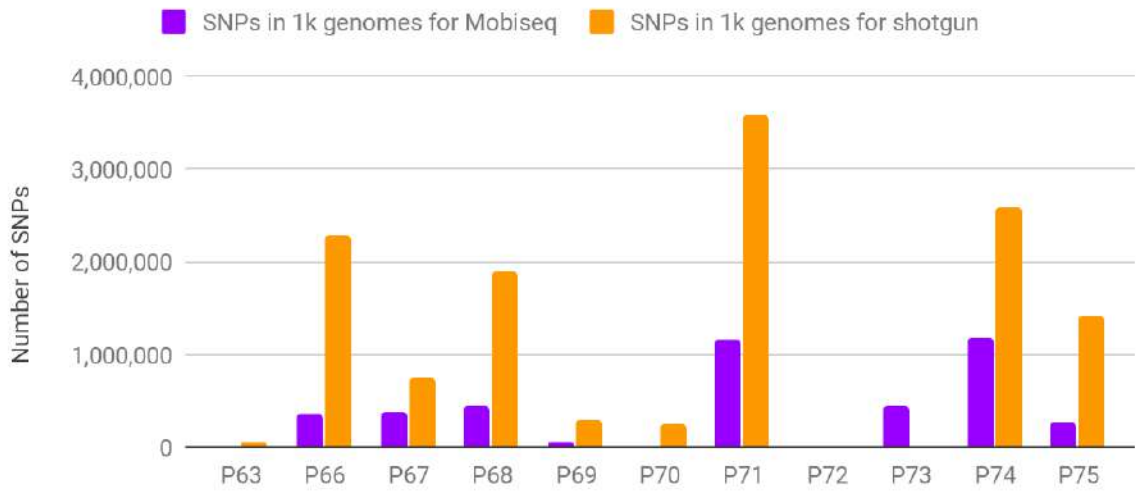


Figure 6. 18. SNPs recovery in 1k genomes for MobiSeq (Alu_v2) and shotgun libraries.

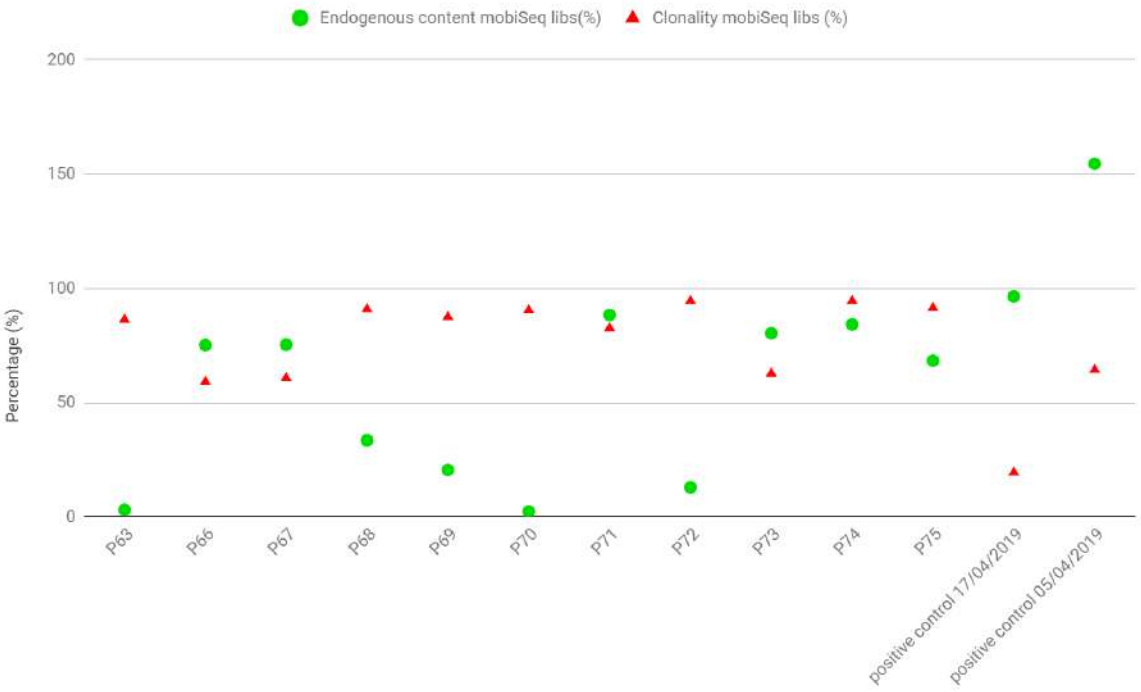


Figure 6. 19. Endogenous content (%) and clonality (%) for MobiSeq Alu_v2 TE-target primer libraries.

7. DISCUSSION AND CONCLUSIONS

7. 1. Mapped reads, endogenous DNA enrichment and clonality

The number of total reads for modern samples is increased compared to the ancient samples. The concentration of the samples was not accurately quantified due to long DNA fragments presented in modern samples. This fact caused an unbalanced titration in the sequence pool, generating high amount of sequenced data that took over parts of ancient DNA sequenced data. An optimal solution for this issue would be a different method of quantification of the positive control concentration.

In addition, to recover more reads including the TE-target primer sequencing mode can be changed to either sequence in pair-end read mode or sequence more bases in single-end read mode (eg. 100 bases). In this study we used Illumina HiSeq 80 bases single-end read mode, thus we might have sequenced only the flanking region instead of both the flanking region and the TE-target primer. Therefore, it would be optimal to map all the filtered reads to the human reference genome for both Alu TE-target primers (not only reads with the Alu TE-target primer).

In single tube libraries, two of the samples (P72 and P73) yielded extremely low coverage of DNA sequenced data due to a mix of indexing primers. That is a common problem in Illumina sequencing platforms (Costello et al. 2018) and further computational analysis might generate more sequence data for those two samples.

The clonality on this study is expected to be high for ancient samples compared to the modern ones. The quality of the starting material is a crucial factor in PCR-based enrichment protocols. Highly degraded samples, such as archaeological speciments, require a high number of PCR cycles for DNA recovery (Avila-Arcos et al. 2011). We can form the hypothesis that the clonality of this method is high and varies according to the quality of the sample. Although, in MobiSeq, the random shearing of DNA prior to library build and the TE-target PCR set up generates fragments with different sequences in the 3’ end of the flanking region. This allows identification of putative PCR duplicates, based on the assumption that any read with no TE-target primer is PCR duplicate.

Endogenous DNA is high in most of the libraries (MobiSeq with both Alu TE-target primers and shotgun libraries). All library building techniques showed >61% average percentage of endogenous content. Increased numbers of endogenous are expected because all the samples were petrous bones. Usually the DNA is well preserved in this part of the skull (Margaryan et al. 2018).

7.2. Troubleshooting

Alu genes present high GC-content which creates biases on the NGS data (Chen et al. 2013). Alu_v1 TE-primer presented a better performance in MobiSeq libraries thus maybe because of the primer length and GC-content. Alu_v1 is 24bp and Alu_v2 is only 18bp. There is also lower GC- content (67%) in Alu_v1 TE-target primer than in Alu_v2 (83%). The primer selection needs to be optimized, probably by using new TE-target primers for different highly abundant transposable elements in the human genome.

Looking into positive control generated data there is an increased number of total reads and high clonality. PCR- based methods usually give lots of duplicates and not primary reads but in modern DNA samples the PCR cycles are lower than the ancient ones, thus the expected clonality tends to be low. Due to the high molecular weight DNA fragments accurate measurement of library concentration cannot be performed using TapeStation High Sensitivity DNA 2200 (Agilent). Non accurate concentration determination creates titration issues in sequencing generating massive number of reads for the same individual. We obtained long fragments of DNA in MobiSeq libraries targeted with both Alu TE-target primers that are chimeric products of the PCR amplification. In addition, in some samples a shorter DNA fragment peak appeared at approximately 80bp, those fragments are PCR duplicates.

Also, consistent with the expectation for degraded DNA, we observed a high level of C → T transitions at the sequence 5′ end. Although, we weren't able to see the same increase in G → A transitions at the 3′ end, because of the presence of the Alu TE-target primer at the 3′ end of each fragment. This pattern has been shown to be a

result of cytosine deamination of single-stranded overhangs in combination with the 3′–5′ exonuclease activity and 5′–3′ polymerase activity of the T4 DNA Polymerase during end repair (Briggs et al. 2007; Brotherton et al. 2007).

7.3. Authenticity of MobiSeq and shotgun libraries

The authenticity of MobiSeq libraries was evaluated by tracking post- mortem damage (deamination, oxidation, depurination) and length distribution per individual (Jónsson et al. 2013). Computational analysis was performed on the unique mapped reads. A weird profile is obtained in both graphs on the 3' end of each fragment including the positive controls, due to the existence of the TE-target primer at the 3' end of each DNA fragment. This is also identified by the length plots of all the samples (short DNA fragments). Also, comparing the positive control damage profile to the MobiSeq libraries, they showed different patterns for each one of the Alu TE-target primers, which confirms that the targeted DNA fragments captured with MobiSeq target capture method is ancient human DNA and there is no contamination of modern DNA (detailed mapDamage profiles for both positive controls in Appendix S4, MapDamage). The high background noise is probably caused by the limited quantity of sequenced DNA data.

The authenticity of shotgun libraries was evaluated using the same pipeline as in MobiSeq libraries. In shotgun libraries the 3' end damage pattern is noticeable because of the absence of any primer. Only mapped reads were processed for mapDamage analysis in shotgun libraries.

7.4. SNPs in 1k genomes in MobiSeq and Shotgun libraries

As described in section 5. 4. 4, a different computational approach than the one described in Rey-Iglesia et al. 2018 was performed in MobiSeq libraries. This method generated more accurate set of results including SNPs number in 1k genomes. We mapped all the reads including those with no TE-target primer to the human reference genome; the analysis that searched for SNPs only in reads with the TE primer was not efficient in this study. Most of the reads containing the TE-target primer were discarded because of the sequencing read mode. As expected we

retrieved higher number of SNPs and more endogenous DNA was recovered, but we still get increased number of clonality. Consequently, further research on optimising this method and statistics analysis is required to have an accurate estimation of the performance of MobiSeq target enrichment approach in ancient human samples.

Finally, no assumptions can be made for the efficiency of MobiSeq method or the TE primers, consequently, an extended statistical analysis is crucial to fully evaluate this capture technique.

8.FUTURE PERSPECTIVES

8.1. TE-target primer design

As mentioned before, Alu genes consist lot of G and C bases which creates biases on the NGS data (Chen et al. 2013). The primer selection needs to be optimized, probably by using new target primers for different highly abundant transposable elements in the human genome.

8.2. Reduce clonality

PCR- based methods usually give lots of duplicates and not primary reads (Aird et al. 2011), which is defined as clonality. Looking into the TapeStation profile of TE-PCR products, we observed that no long fragments of DNA are presented, or their amount is really low. However, after index PCR the concentration of those long fragments is dramatically increased. Quantifying the TE-PCR products might reduce the index PCR cycles required for an optimal concentration of the final (indexed) PCR product, which will reduce the clonality.

8.3. Titration

Because of high molecular weight DNA fragments quantification of library concentration was not accurate creating titration issues in sequencing. This generated massive number of reads for the same individual and the positive controls that took over sequenced reads of luckily informative data. Other methods for measuring DNA concentration could be used (eg. Qubit dsDNA High Sensitivity (HS) assay (Life technologies) or TapeStation Genomic DNA (Agilent)). Another solution would be the quantification of the TE-PCR products before index PCR performance which will reduce the longer DNA fragments appeared in the final (indexed) PCR product.

8.4. Sequencing read mode and bioinformatic process

Previous studies on MobiSeq target enrichment method have accomplished to generate 90% of loci across the genome, and performed SNP discovery with relatively low rates of clonality in non- human species (Rey-Iglesia et al. 2018).

Comparing our results to this study we came to the point that sequencing read mode was not efficient for this method. A false perspective was that after adapter removal we targeted only the reads that contained the Alu TE-target primer so we might lose lots of variable reads. Nevertheless, we used Illumina HiSeq single-end read mode 80 bases sequencing thus we might not sequence the TE-target primer, but only the flanking region. To recover more reads including the TE-target primer we could change the sequencing read mode, either by sequence in pair-end read mode or sequence more bases in single read mode (eg. 100 bases), which is more expensive.

Downstream analysis was performed again using all the unique reads (including and excluding TE primer). As expected, we retrieved higher number of SNPs and more endogenous DNA was recovered, but we get increased number of clonality. Consequently, further research on optimising this methods and statistics analysis is required to have an accurate estimation of the performance of MobiSeq target enrichment approach in ancient human samples.

9. REFERENCES

10. Adler CJ, Haak W, Donlon D, Cooper A, Consortium TG. (2011). Survival and recovery of DNA from ancient teeth and bones. *Journal of Archaeological Science, 38(5)*, 956- 964.
11. Ali OA, O'Rourke SM, Amish SJ, Meek MH, Luikart G, Jeffres C, & Miller MR. (2016). RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics, 202(2),* 389-400.
12. Allentoft ME, Sikora M, Sjögren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L, Malaspinas AS, Margaryan A, Higham T, Chivall D, Lynnerup N, Harvig L, Baron J, Della Casa P, Dąbrowski P, Duffy PR, Ebel AV, Epimakhov A, Frei K, Furmanek M, Gralak T, Gromov A, Gronkiewicz S, Grupe G, Hajdu T, Jarysz R, Khartanovich V, Khokhlov A, Kiss V, Kolář J, Kriiska A, Lasak I, Longhi C, McGlynn G, Merkevicius A, Merkyte I, Metspalu M, Mkrtchyan R, Moiseyev V, Paja L, Pálfi G, Pokutta D, Pospieszny Ł, Price TD, Saag L, Sablin M, Shishlina N, Smrčka V, Soenov VI, Szeverényi V, Tóth G, Trifanova SV, Varul L, Vicze M, Yepiskoposyan L, Zhitenev V, Orlando L, Sicheritz-Pontén T, Brunak S, Nielsen R, Kristiansen K, Willerslev E. (2015). Population genomics of Bronze Age Eurasia. *Nature, 522(7555),* 167- 172.
13. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. (1981). Sequence and organization of the human mitochondrial genome. *Nature,* 290(5806) 457-465.
14. Andrews KR, Good JM, Miller MR, Luikart G, & Hohenlohe PA. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics, 17(2),* 81.
15. Andrews KR, Hohenlohe PA, Miller MR, Hand BK, Seeb JE, & Luikart G. (2014). Trade-offs and utility of alternative RADseq methods: Reply to Puritz et al. *Molecular ecology, 23(24),* 5943-5946.
16. Andrews KR, & Luikart G. (2014). Recent novel approaches for population genomics data analysis. *Molecular ecology, 23(7)*, 1661-1667.
17. Avila-Arcos MC, Cappellini E, J. Romero-Navarro A, Wales N, Moreno-Mayar JV, Rasmussen M, Fordyce SL, et al. (2011). Application and Comparison of

Large-Scale Solution-Based DNA Capture-Enrichment Methods on Ancient DNA. *Scientific Reports 1*, 74.

18. Ávila-Arcos MC, Sandoval-Velasco M, Schroeder H, Carpenter M. L, Malaspinas AS, Wales N, Peñaloza F, Bustamante C. D, Gilbert, MT. (2015). Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries. *Methods Ecol Evol* 6(6), 725–734.
19. Briggs AW, Heyn P. (2012). Preparation of next-generation sequencing libraries from damaged DNA. *Methods Mol Biol, 840,* 143-154.
20. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, & Johnson EA. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one, 3(10),* e3376.
21. Barreiro FS, Vieira FG, Martin DM, Haile J, Gilbert T, and Wales N. 2017. “Characterizing Restriction Enzyme-Associated Loci in Historic Ragweed (Ambrosia Artemisiifolia) Voucher Specimens Using Custom-Designed RNA Probes.” *Molecular Ecology Resources*. http://dx.doi.org/10.1111/1755-0998.12610
22. Batzer MA, Deininger PL. (2002). Alu repeats and human genomic diversity. *Nat Rev Genet, 3(5),* 370-379.
23. Blow MJ, Zhang T, Woyke T, Speller CF, Krivoshapkin A, Yang DY, Derevianko A, Rubin EM. (2008). Identification of ancient remains through genomic sequencing, *Genome Res, 18(8),* 1347-1353.
24. Bodi, K, Perera, A. G., Adams, P. S., Bintzler, D., Dewar, K., Grove, D. S., Zianni, M. (2013). Comparison of Commercially Available Target Enrichment Methods for Next-Generation Sequencing, *Journal of Biomolecular Techniques 24*, 73- 86.
25. Bos KI, Jager G, Schuenemann VJ, Vagene AJ, Spyrou MA, Herbig A, Nieselt K, Krause. (2014). Parallel detection of ancient pathogens via array-based DNA capture, *Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1660)* https://doi.org/10.1098/rstb.2013.0375
26. Briggs AW, and Heyn P. (2012). Preparation of next-Generation Sequencing Libraries from Damaged DNA. *Methods in Molecular Biology 840*, 143–54.
27. Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. (2007). Patterns of damage in

genomic DNA sequences from a Neandertal, *Proceedings of the National Academy of Sciences of the United States of America, 104(37),* 14616- 21.

28. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, Austin J, Cooper A. (2007). Novel high-resolution characterization of ancient DNA reveals C>U-type base modification events as the sole cause of post mortem miscoding lesions, *Nucleic Acids Res, 35(17)*, 5717-5728.
29. Burbano HA, Green RE, Maricic T, Lalueza-Fox C, de la Rasilla M, Rosas A, Kelso J, Pollard KS, Lachmann M, Pääbo S. (2012). Analysis of human accelerated DNA regions using archaic hominin genomes. *PLoS One 7 (3),* e32877.
30. Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PL, Xuan Z, Rooks M, Bhattacharjee A, Brizuela L, Albert FW, de la Rasilla M, Fortea J, Rosas A, Lachmann M, Hannon GJ, Pääbo S. (2010). Targeted investigation of the Neandertal genome by array based sequence capture. *Science 328(5979)*, 723–725.
31. Cabana, Graciela S., Brannon I. Hulsey, and Frankie L. Pack. (2013). Molecular Methods. *Research Methods in Human Skeletal Biology*. http://dx.doi.org/10.1016/b978-0-12-385189-5.00016-9
32. Carøe C, Gopalakrishnan S, Vinner L, Mak S, Sinding M, Samaniego J, Wales N, Sicheritz- Pontén T, and Gilbert T. (2018). Single- tube Library Preparation for Degraded DNA. *Methods in Ecology and Evolution / British Ecological Society*. http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12871/full
33. Carpenter L.M, Buenrostro D. J, Valdiosera C, Schroeder H, Allentoft E. M, Sikora M, Rasmussen M, Gravel S, Guillén S, Nekhrizov G, Leshtakov K, Dimitrova D, Theodossiev N, Pettener D, Luiselli D, Sandoval K, Moreno-Estrada A, Li Y, Wang J, Gilbert M. T. , Willerslev E, Greenleaf J. W, Bustamante D. C. (2013). Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *Am J Hum Genet 93(5)*, 852- 864.
34. Carpenter S, Aiello D, Atianand MK, Ricci EP, Gandhi P, Hall LL, Byron M, et al. (2013). A Long Noncoding RNA Mediates Both Activation and Repression of Immune Response Genes. *Science* 341 (6147): 789–92.

35. Chen C, Ara T, Gautheret D. (2009). Using Alu elements as polyadenylation sites: A case of retroposon exaptation. *Mol Biol Evol*, *26,* 327- 334.
36. Chen C, Gentles AJ, Jurka J, & Karlin S. (2002). Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA 99*, 2930- 2935.
37. Chénais B, Caruso ., Hiard S, & Casse N. (2012). The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene, 509(1)*, 7-15.
38. Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ (1999). Genome-wide mapping with biallelic markers in Arabidopsis thaliana. *Nat. Genet. 23(2)*, 203-207.
39. Cordaux R, Batzer M.A. (2009). The impact of retrotransposons on human genome evolution, *Nat. Rev. Genet. 10(10),* 691- 703.
40. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, et al. (2018). Characterization and Remediation of Sample Index Swaps by Non-Redundant Dual Indexing on Massively Parallel Sequencing Platforms. *BMC Genomics* 19 (1), 332.
41. Coyne KJ, Burkholder J, Feldman, RA, Hutchins DA, & Cary SC. (2004). Modified serial analysis of gene expression method for construction of gene expression profiles of microbial eukaryotic species. *Applied and environmental microbiology, 70(9),* 5298-5304.
42. Cruz-Dávalos DI, Llamas B, Gaunitz C, Fages A, Gamba C, Soubrier J, Orlando L. (2017). Experimental conditions improving in-solution target enrichment for ancient DNA. *Molecular Ecology Resources, 17(3),* 508- 522.
43. ———. 2017b. “Experimental Conditions Improving in-Solution Target Enrichment for Ancient DNA.” *Molecular Ecology Resources* 17 (3): 508–22.
44. Damgaard PB, et al. (2015) Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep* 5:11184.
45. Daniel C, Silberberg G, Behm M, Ohman M. (2014). Alu elements shape the primate transcriptome by cis-regulation of rna editing, *Genome Biol. 15*, R28.
46. Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics, 9(5-6),* 416- 423.

47. Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. (2013). Special features of RAD Sequencing data: implications for genotyping. *Mol Ecol, 22(11),* 3151-3164.
48. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, & Blaxter ML. (2011). Genome-wide genetic marker discovery and genotyping using next generation sequencing. *Nature Reviews Genetics, 12(7),* https://doi.org/10.1038/nrg3012
49. Deng J, Shoemaker R, Xie B, Gore A, LeProust EM, Antosiewicz-Bourget J, Egli D, Maherali N, Park IH, Yu J, Daley GQ, Eggan K, Hochedlinger K, Thomson J, Wang W, Gao Y, Zhang K. (2009). Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. *Nat Biotechnol, 27 (4),* 353-360.
50. Devault AM, McLoughlin K, Jaing C, Gardner S, Porter TM, Enk JM, Thissen J, Allen J, Borucki M, DeWitte SN, Dhody AN, Poinar HN. (2014). Ancient pathogen DNA in archaeological samples detected with a Microbial Detection Array, *Sci Rep, 4,* 4245.
51. Dewannieux M, Esnault C, Heidmann T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet*, *35*, 41- 48.
52. Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, Chanrion B, Fu Y, Su T, Hibshoosh H. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res. 20(7)*, 972- 980.
53. Elshire RJ, Glaubitz JC, Sun Q, Poland JA., Kawamoto K, Buckler ES, & Mitchell SE. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one, 6(5)*, e19379.
54. Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE, & Holzapfel CM. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the national academy of sciences, 107(37)*, 16196-16200.
55. Etter PD, Preston JL, Bassham S, Cresko WA, & Johnson EA. (2011). Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PloS one, 6(4),* e18561.
56. Federhen S. (2012). The NCBI taxonomy database. *Nucleic Acids Res. 40(D1)*, D136–D143.

57. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, Viola B, Prüfer K, Meyer M, Kelso J, Reich D, Pääbo S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature, 524(7564),* 216-219.
58. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S. (2013). DNA analysis of an early modern human from Tianyuan Cave, China, *Proc Natl Acad Sci USA, 110(6),* 2223- 2227.
59. Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations. *Molecular ecology 26(20),* 5369-5406.
60. Gaither MR, Bernal MA, Coleman RR, Bowen BW, Jones SA., Simison WB, & Rocha LA. (2015). Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology, 24(7)*, 1543-1557.
61. Gansauge MT, Meye, M. (2014). Selective enrichment of damaged DNA molecules for ancient genome sequencing, *Genome Research, 24 (9),* 1543-9.
62. Gardner S, Jaing C, McLoughlin K, & Slezak T. (2010). A microbial detection array (MDA) for viral and bacterial detection. *BMC Genomics 11*, 668.
63. Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology, 22(11)*, 3165-3178.
64. Ginolhac G, Forster P, Pascal F, Ovarlez J. (2012). Derivation of the Bias of the Normalized Sample Covariance Matrix in a Heterogeneous Noise With Application to Low Rank STAP Filter. *IEEE Trans Signal Process, 60(1),* 514-518.
65. Glocke I, Meyer M. (2017). Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. *Genome Res*:gr.219675.116.
66. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol 27*, 182–189.
67. Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, Gogol-Döring A, Kapitonov V, Diem T, Dalda A, Jurka J, Pritham EJ, Dyda F, Izsvák

Z, Ivics Z. (2016). A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nature communications, 7*, 10716.

68. Graham CF, Glenn TC, McArthur AG., Boreham DR, Kieran T, Lance S et al. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). *Molecular Ecology Resources, 15(6)*, 1304-1315.
69. Graham E. (2007). DNA reviews: Ancient DNA. *Forensic Science, Medicine, and Pathology*, *3 (3)*, 221–225.
70. Gray MW. (2001). Mitochondrial DNA (mtDNA). New York, Academic Press, ISBN 9780122270802, *Encyclopedia of Genetics,* 1219-1220.
71. Gray, Y. H. (2000). It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics, 16(10),* 461-468.
72. Green RE, Malaspinas AS, Krause J, Briggs AW, et al. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell,* ISBN 0092-8674.
73. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Bánffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O,Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szécsényi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D. (2015). Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature, 522(7555),* 207-211.
74. Hansen HB, Damgaard PB, Margaryan A, Stenderup J, Lynnerup N, Willerslev E, Allentoft ME. (2017). Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One 12(1),* e0170940.
75. Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, Brundage T, Boudreau A, Chow S, Eberle J, Erbilgin A, Falkowski M, Fitzgerald R, Ghose S, Iartchouk O, Jain M, Karlin-Neumann G, Lu X, Miao X, Moore B, Moorhead M, Namsaraev E, Pasternak S, Prakash E, Tran K, Wang Z, Jones HB, Davis RW, Willis TD, Gibbs RA (2005). Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res, 15(2)*, 269-275.

76. Hohenlohe PA., Bassham S, Etter PD, Stiffler N, Johnson EA, & Cresko WA. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics, 6(2)*, e1000862.
77. Huang H, & Knowles LL. (2014). Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic biology, 65(3)*, 357-365.
78. Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics,* 29(13), 1682-1684.
79. Jurka J. (2000). Repbase update a database and an electronic journal of repetitive elements. *Trends Genet. 16,* 418- 420.
80. Jurka J, Kapitonov V.V, Kohany O, Jurka M.V. (2007). Repetitive sequences in complex genomes: structure and evolution, *Annu. Rev. Genom. Hum. Genet., 8,* 241- 259.
81. Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. *Science, 303 (5664)*, 626- 1632.
82. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. (2002). The human genome browser at UCSC, *Genome Res, 12 6),* 996-1006.
83. Kivisild T. (2017). The study of human Y chromosome variation through ancient DNA, *Hum Genet, 136(5),* 529-546.
84. Knapp M, Clarke AC, Ann Horsburgh K, Matisoo-Smith EA. (2012). Setting the stage – Building and working in an ancient DNA laboratory. *Annals of Anatomy - Anatomischer Anzeiger, 194(1),* 3-6.
85. Knapp M, Hofreiter M. (2010). Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes, 1(2),* 227-243.
86. Korenberg, J. R. & Rykowski, M. C. (1988). Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell 53*, 391- 400.
87. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al. (2001). Initial sequencing and analysis of the

human genome. International Human Genome Sequencing Consortium. *Nature*, *409,* 860- 921.

88. Li H, Durbin R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26(5),* 589-595.
89. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25(16),* 2078-2079.
90. Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. (1998). Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat Genet, 19(3),* 225-232.
91. Lovett, M., Kere, J. & Hinton, L.M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA 88(21),* 9628-9632.
92. Luan DD, Korman MH, Jakubczak JL, & Eickbush TH. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non- LTR retrotransposition. *Cell, 72(4)*, 595- 605.
93. Luikart G, England PR, Tallmon D, Jordan S, & Taberlet P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature reviews genetics, 4(12),* 981.
94. Luikart G, Ryman N, Tallmon DA, Schwartz MK, Allendorf FW. (2010). Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches. *Conserv Genet, 11(2)*, 355-373.
95. Mackelprang R, Rubin EM. (2008). Paleontology. New tricks with old bones. *Science, 321 (5886)*, 211-212.
96. Mak, Sarah Siu Tze, Shyam Gopalakrishnan, Christian Carøe, Chunyu Geng, Shanlin Liu, Mikkel-Holger S. Sinding, Lukas F. K. Kuderna, et al. 2017. “Comparative Performance of the BGISEQ-500 vs Illumina HiSeq2500 Sequencing Platforms for Palaeogenomic Sequencing.” *GigaScience*. http://dx.doi.org/10.1126/science.1161890
97. Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Turner, D. J. (2010). Target- enrichment strategies for next- generation sequencing, *Nature Methods 7(2),* 111-118.

98. Margaryan A, Hansen HB, Rasmussen S, Sikora M, Moiseyev V, Khoklov A, Epimakhov A, Yepiskoposyan L, Kriiska A, Varul L, Saag L, Lynnerup N, Willerslev E, Allentoft ME. (2018). Ancient pathogen DNA in human teeth and petrous bones. *Ecol Evol,* 8(6), 3534-3542.
99. Maricic T, Whitten M, Pääbo S. (2010). Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One, 5(11),* e14004
100. Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal* 17 (1): 10–12.
101. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, de Castro JM, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MA, Roodenberg J, Vergès JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature, 528(7583),* 499-503.
102. Meyer M., & Kircher M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *6*, pdb-prot5448.
103. Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. (2007). Targeted high- throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res, 35(15)*, e97.
104. Mourier T., & Willerslev E. (2009). Retrotransposons and non-protein coding RNAs. *Briefings in Functional Genomics and Proteomics, 8(6)*, 493-501.
105. Ozga AT, Nieves-Colón MA, Honap TP, Sankaranarayanan K, Hofman CA, Milner GR, Lewis Jr CM, Stone AC, Warinner C. (2016). Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus, *Am J Phys Anthropol, 160,* 220- 228.
106. Pääbo S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci U S A, 86(6),* 1939-1943.

107. Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. (2004). Genetic Analyses from Ancient DNA, *Annual Review of Genetics*, *38(1),* 645- 679.

108. Paijmans JL, Fickel J, Courtiol A, Hofreiter M, Förster DW, (2016). Impact of enrichment conditions on cross- species capture of fresh and degraded DNA. *Mol Ecol Resour, 16 (1),* 42- 55.

109. Pakendorf B, Stoneking M. (2005). Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet 6:* 165–183.

110. Parimoo S., Patanjali SR., Shukla H., Chaplin DD. & Weissman SM. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA 88(21),* 9623-9627.

111. Peterson BK., Weber JN., Kay EH., Fisher HS., & Hoekstra HE. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one, 7(5),* e37135.

112. Platt RN., Vandewege MW., & Ray DA. (2018). Mammalian transposable elements and their impacts on genome evolution. *Chromosome Research,* 1-19.

113. Quinlan, A. R., & Hall, I. M. 2010. “BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics.” *Bioinformatics* 26(6), 841-842.

114. Renaud G, Slon V, Duggan AT, Kelso J. (2015). Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol, 16*, 224.

115. Rey-Iglesia A, Gopalakrishnan S, Carøe C et al. (2018) MobiSeq: De Novo SNP discovery in model and non- model species through sequencing the flanking region of transposable elements. *bioRxiv*. Available at: https://www.biorxiv.org/content/early/2018/06/17/349290.abstract

116. Rohland N, Glocke I, Aximu-Petri A, Meyer M. (2018). Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat Protoc, 13(11)*, 2447- 2461.

117. Rohland N, Hofreiter M. (2007). Comparison and optimization of ancient DNA extraction. *Biotechniques 42(3)*, 343- 352.

118. Rowold D.J, Herrera R.J. (2000). Alu elements and the human genome, *Genetica Miami 33199, USA, 108(1)*, 57- 72.

119. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science, 239(4839)*, 87-49. **This paper was the first description of PCR, which, coupled to electrophoretic sequencing, is the primary conventional method for targeted variation analysis.**

120. Schmid S, Genevest R, Gobet E, Suchan T, Sperisen C, Tinner W, & Alvarez N. (2017). HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. *Methods in Ecology and Evolution, 8(10)*, 1374-1388.

121. Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, and Orlando L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc., 9(5)*, 1056-82.

122. Schubert, Lindgreen, and Orlando (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes, 9(1),* 88.

123. Schweyen H, Rozenberg A, & Leese F. (2014). Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *The Biological Bulletin, 227(2)*, 146-160.

124. Sela N, Mersch B, Hotz-Wagenblatt A, & Ast G. (2010). Characteristics of transposable element exonization within human and mouse. *PLoS One, 5(6)*, e10907.

125. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A, 111(6)*, 2229- 2234.

126. Smit AFA. and Riggs AD. (1996) Tiggers and other DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA 93*, 1443- 1448.

127. Wagner CE., Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O (2013). Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Molecular ecology, 22(3), 787-798.

128. Wang DG, (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science, 280(5366)*, 1077-1082.

129. Willerslev E, Cooper A. (2005). Ancient DNA. *Proc Biol Sci, 272 (1558),* 3-16.

Appendix S1- Figures and Tables

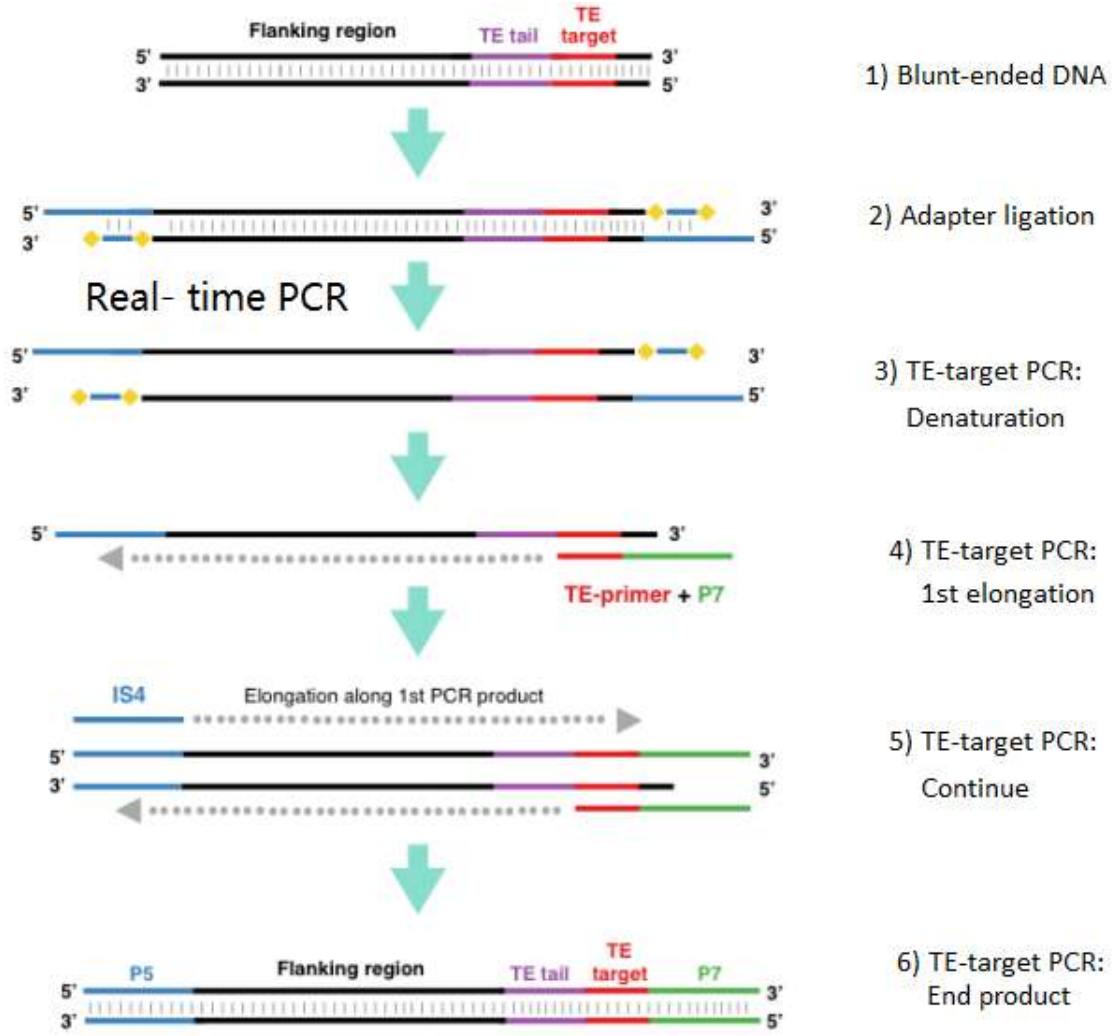


Figure S1. Schematic overview of the MobiSeq protocol. 1) Blunt-end repair, overhanging 5' and 3' ends are filled in or removed by T4 DNA polymerase. 5' phosphates are attached using T4 polynucleotide kinase. 2) Double-stranded mP5 adapters are ligated to the fragment by T4 DNA Ligase. Adapters do not carry 5' phosphates and therefore only one strand is ligated to the fragments. 3) TE-target PCR is performed using forward IS4 primer and the TE-fusion primer. TE-target PCR will also result in the incorporation of P7 sequencing adapter. 4) PCR elongation will only occur upstream of the 3' end of the TE-target primer. 5) The end product of the TE-target PCR will be the TE-target sequence, a TE tail and the flanking genomic region. It will also contain Illumina sequencing adapters (P5 and P7). This PCR product can be indexed using single or dual indexing primers, then pooled for sequencing with other samples. Yellow diamonds represent mP5 adapter modifications (adapted from (Rey-Iglesia, Gopalakrishnan, and Carøe 2018)).

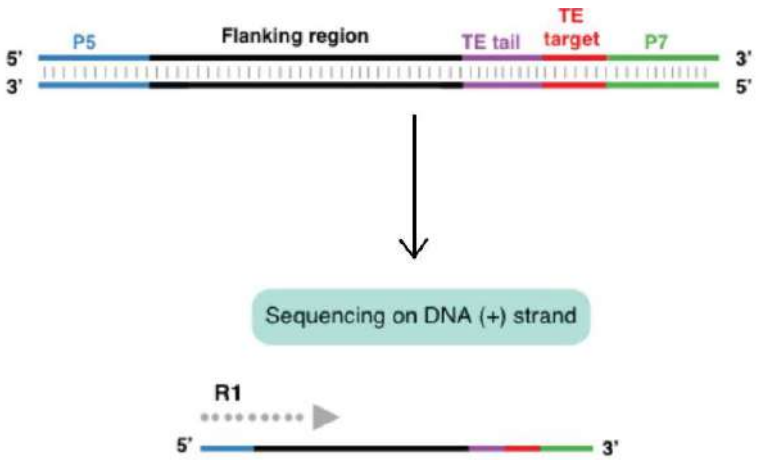


Figure S2. Sequencing architecture of Transposable Element (TE) flanking region in the (+) strand. The TE-target primer will always be at the 3'end of Read 1 (R1). In this occasion "wrong" read cannot be sequenced here, since the fragment is short and most likely we are going to cover the whole fragment.

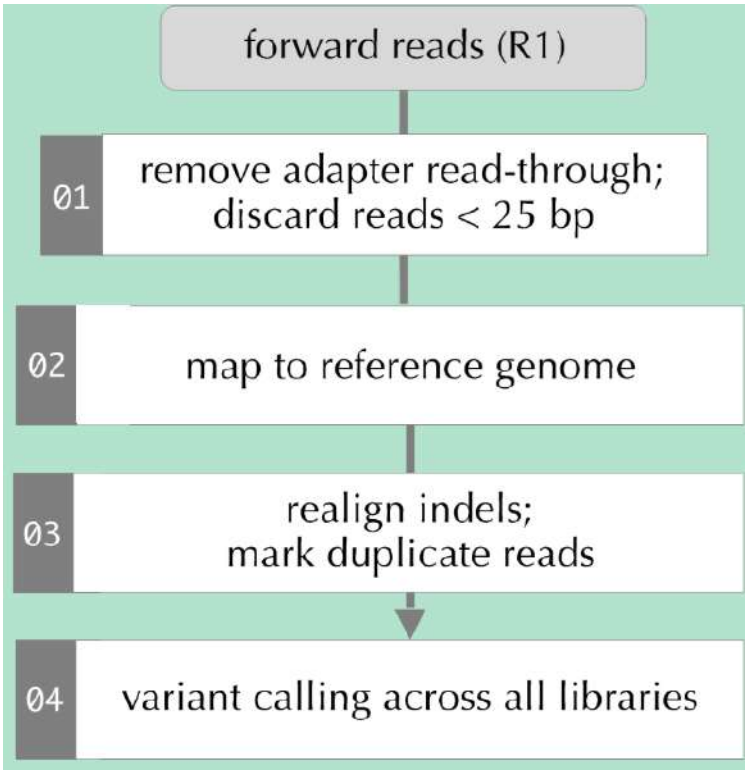


Figure S3. Schematic overview of the data processing pipeline used in this study.

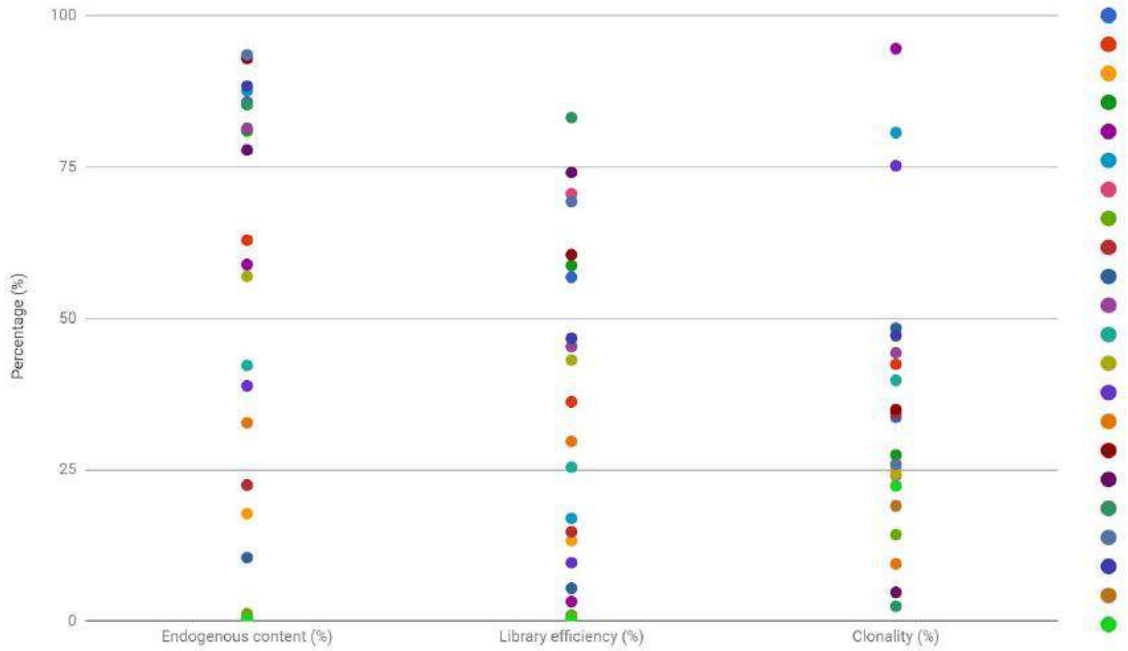


Figure S4. Endogenous content (%), library efficiency (%) and clonality (%) deviation across the samples for shotgun libraries.

Table S1. Sample information.

Sample name	Sample ID	Excavation ID	Region	Bone	Bone Side
P56	P56/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P57	P57/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P58	P58/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P59	P59/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P60	P60/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P61	P61/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P62	P62/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P63	P63/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P64	P64/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P65	P65/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P66	P66/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P67	P67/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P68	P68/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P69	P69/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P70	P70/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P71	P71/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P72	P72/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P73	P73/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P74	P74/ 2019 KMG	KBM 4074	Denmark	petrous bone	right
P75	P75/ 2019 KMG	KBM 4074	Denmark	petrous bone	right

Table S2. Overview of the oligonucleotides used in this study.

Primer name	Description	Sequence	Annealing (°C) in target PCR	Reference
ALU_v1	TE-target+P7	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGATTACAGGCGTGAGCCACCGCGCC-3	65°C	This study
ALU_v2	TE-target+P7	GCTCTTCCGATCTCATGAGCCACCGCGCCCGGC-3'	65°C	This study
IS1	For mP5 adaptor	5'-A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T-3'		Meyer & Kircher (2010)
modifiedIS3	For mP5 adaptor	5'-(no 5' PHO)A*G*A*T*CGGAA*G*A*G*C*[C3 spacer]-3'		Carøe et al. (2017)
IS7	Short forward primer Target-PCR	5'-ACACTCTTTCCCTACACGAC-3'		Meyer & Kircher (2010)
IS4	Short forward primer Target-PCR / Forward primer for index PCR	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT-3'		Meyer & Kircher (2010)
P7	Reverse index primer	5'-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGT-3'		Meyer & Kircher (2010)

Table S3a. Real time PCR results for MobiSeq libraries.

Primer name	Sample ID	Ct/ Cp values	Cycles
Alu v1	P56/ 2019 KMG	28.54	30
	P57/ 2019 KMG	27.37	28
	P58/ 2019 KMG	26.84	28
	P59/ 2019 KMG	25.36	26
	P60/ 2019 KMG	29.17	30
	P61/ 2019 KMG	26.67	28
	P62/ 2019 KMG	21.39	22
	P63/ 2019 KMG	27.94	29
	P64/ 2019 KMG	22.6	24
	P65/ 2019 KMG	24.95	26
	positive control 17/04/2019	18.92	20
	positive control 5/04/2019	-	20
Alu v2	P66/ 2019 KMG	26.21	27
	P67/ 2019 KMG	26.57	28
	P68/ 2019 KMG	25.48	26
	P69/ 2019 KMG	26.47	27
	P70/ 2019 KMG	26.08	27
	P71/ 2019 KMG	25.32	26
	P72/ 2019 KMG	26	27
	P73/ 2019 KMG	-	20
	P74/ 2019 KMG	26.84	28
	P75/ 2019 KMG	26.5	28
	positive control 17/04/2019	20.75	22
	positive control 5/04/2019	-	20

Table S3b. Real time PCR results for Single-tube libraries.

Sample ID	Ct/ Cp values	Cycles
P56/ 2019 KMG	21.28	22
P57/ 2019 KMG	21.31	22
P58/ 2019 KMG	19.89	21
P59/ 2019 KMG	20.34	21
P60/ 2019 KMG	29.2	30
P61/ 2019 KMG	26.39	27
P62/ 2019 KMG	20.11	21
P63/ 2019 KMG	17.58	19
P64/ 2019 KMG	20.75	22
P65/ 2019 KMG	22.18	23
P66/ 2019 KMG	21.41	22
P67/ 2019 KMG	21.37	22
P68/ 2019 KMG	19.72	21
P69/ 2019 KMG	24.1	25
P70/ 2019 KMG	17.09	18
P71/ 2019 KMG	20.7	22
P72/ 2019 KMG	22.77	24
P73/ 2019 KMG	18.12	19
P74/ 2019 KMG	20.36	21
P75/ 2019 KMG	22.01	23

Table S4 Average sequencing and mapping statistics per sample for both TE-target primers.

Sample name	Alu primer	Total reads	Reads after adapter removal	hg19 mapped primer	Unique reads	Endogenous content (%)	Clonality (%)
P56	Alu_v1	1,386,356	1,334,934	1,175,324	21,428	85	98
P57	Alu_v1	4,546,981	4,416,869	3,813,535	288,775	84	92
P58	Alu_v1	7,706,551	7,453,236	6,007,006	515,278	78	91
P59	Alu_v1	8,626,208	8,290,590	7,489,683	907,096	87	88
P60	Alu_v1	3,743,659	3,604,917	3,227,991	86,165	86	97
P61	Alu_v1	6,076,376	5,967,000	5,628,107	418,588	93	93
P62	Alu_v1	6,273,043	6,088,360	5,718,411	3,029,508	91	47
P63	Alu_v1	2,235,394	2,196,890	363,827	24,796	16	93
P64	Alu_v1	1,579,602	1,513,474	709,485	117,879	45	83
P65	Alu_v1	1,108,232	1,039,750	523,102	66,606	47	87
P73	Alu_v1	1,497,297	1,450,721	1,328,775	641,076	89	52
P63	Alu_v2	5,758,510	3,947,839	3,947,839	177,119	3	86
P66	Alu_v2	835,827	746,015	746,015	628,092	75	59
P67	Alu_v2	901,763	805,315	805,315	679,628	75	61
P68	Alu_v2	15,438,996	13,500,622	13,500,622	5,170,367	33	91
P69	Alu_v2	2,670,040	1,663,019	1,663,019	548,397	21	88
P70	Alu_v2	1,504,469	144,592	144,592	35,657	2	90
P71	Alu_v2	6,978,454	6,710,462	6,710,462	1,073,464	88	83
P72	Alu_v2	1,540,135	262,226	262,226	75,457	13	95
P73	Alu_v2	1,511,487	1,431,396	1,431,396	595,990	80	63
P74	Alu_v2	23,557,045	21,472,991	21,472,991	1,085,452	84	95
P75	Alu_v2	4,291,292	3,443,332	3,443,332	988,250	68	92

Table S5. Average sequencing and mapping statistics per Alu TE-target primer.

Alu TE-target primer	Average of total reads	Average of reads after adapter removal	hg19 mapped primer	Unique reads	Endogenous content (%)	Clonality (%)
Alu_v1	9,359,720	9,153,491	8,461,510	4,059,067	76	76
Alu_v2	9,065,706	8,141,760	9,710,537	3,214,465	61	76

Table S6. Average sequencing and mapping statistics per sample for shotgun libraries.

Sample name	Total reads	Reads after adapter removal	Mapped reads	Unique reads
P56	2,652,217	2,640,506	2,271,786	1,506,143
P57	3,203,499	3,191,339	2,014,331	1,160,134
P58	2,781,971	2,763,987	493,890	369,867
P59	2,599,991	2,597,049	2,104,186	1,526,728
P60	2,320,821	2,135,793	1,366,113	75,001
P61	1,426,793	1,417,909	1,249,351	241,822
P62	3,006,092	3,005,141	2,792,031	2,120,959
P63	3,448,036	3,389,405	41,702	35,742
P64	3,437,479	3,407,420	771,959	506,815
P65	2,830,908	2,809,509	297,429	153,624
P66	3,213,677	3,211,978	2,614,751	1,456,982
P67	2,126,751	2,094,430	897,625	540,630
P68	3,083,097	3,057,652	1,755,280	1,329,224
P69	3,172,824	3,104,167	1,232,196	305,751
P70	2,637,811	2,625,021	863,929	782,517
P71	3,763,571	3,761,820	3,498,441	2,276,582
P72	54	54	42	40
P73	237	237	202	197
P74	2,354,820	2,353,031	2,201,581	1,631,374
P75	2,124,508	2,121,221	1,876,677	991,974
Extraction blank	437,274	39,497	2,084	1,687
Library blank	511,601	36,476	1,717	1,225

Table S7. Number of SNPs per MobiSeq library (Alu_1 TE-target primer) and shotgun libraries.

Sample name/ Alu primer	SNPs in 1k genomes for MobiSeq	SNPs in 1k genomes for shotgun	Average coverage of SNPs for MobiSeq libraries	Average coverage of SNPs for shotgun libraries
P56 Alu_v1	19,307	2,080,845	1.17	1.03
P57 Alu_v1	309,803	90,087	1.27	1.03
P58 Alu_v1	487,888	552,565	1.46	1.02
P59 Alu_v1	767,436	204,240	1.72	1.03
P60 Alu_v1	81,153	31,270	1.19	1.03
P61 Alu_v1	501,934	319,256	1.28	1.02
P62 Alu_v1	1,723,970	3,300,570	3.14	1.03
P63 Alu_v1	15,428	49,808	1.41	1.00
P64 Alu_v1	107,484	633,253	1.24	1.02
P65 Alu_v1	52,634	160,438	1.25	1.02
P73 Alu_v1	660,806	350	1.56	1.00
P63 Alu_v2	10,904	49,808	1.71	1.00
P66 Alu_v2	355,516	2,280,667	1.21	1.03
P67 Alu_v2	367,464	755,048	1.22	1.02
P68 Alu_v2	445,645	1,885,020	1.40	1.03
P69 Alu_v2	59,119	295,919	1.22	1.05
P70 Alu_v2	3,094	242,700	1.09	1.02
P71 Alu_v2	1,149,969	3,591,804	1.54	1.04
P72 Alu_v2	13,226	55	1.10	1.00
P73 Alu_v2	445,185	350	1.58	1.00
P74 Alu_v2	1,187,156	2,583,271	1.44	1.03
P75 Alu_v2	279,516	1,418,161	1.24	1.03

Table S8. Endogenous content (%) and clonality (%) per MobiSeq library.

Sample name/ Alu primer	Endogenous content (%)	Clonality (%)
P56 Alu_v1	85	98
P57 Alu_v1	84	92
P58 Alu_v1	78	91
P59 Alu_v1	87	88
P60 Alu_v1	86	97
P61 Alu_v1	93	93
P62 Alu_v1	91	47
P63 Alu_v1	16	93
P64 Alu_v1	45	83
P65 Alu_v1	47	87
P73 Alu_v1	89	52
positive control 17/04/2019 Alu_v1	97	16.8
positive control 05/04/2019 Alu_v1	96	40.1
P63 Alu_v2	3	86
P66 Alu_v2	75	59
P67 Alu_v2	75	61
P68 Alu_v2	33	91
P69 Alu_v2	21	88
P70 Alu_v2	2	90
P71 Alu_v2	88	83
P72 Alu_v2	13	95
P73 Alu_v2	80	63
P74 Alu_v2	84	95
P75 Alu_v2	68	92
positive control 17/04/2019 Alu_v2	97	19
positive control 05/04/2019 Alu_v2	155	64

Appendix S2- MobiSeq protocol

MobiSeq: De Novo SNP discovery in ancient humans through sequencing the flanking region of transposable elements

Library build - End repair (Blunt-end single-tube library build)

1) Mix the following components in a PCR tube:

Sample input	32 µL
Number of libraries	

End-Repair master mix

Reagent	Stock conc.	Final conc.	Total U	V/R µL	X
T4 DNA polymerase	3 U/µL	0.03	1.2	0.4	
T4 PNK	10 U/µL	0.25	10	1	
dNTPs	25 mM	0.25	-	0.4	
T4 DNA ligase buffer (NEB)	10x	1x	-	4	
Reaction enhancer (see buffer preparation)				2.2	
Total:				**8**	

Reaction size:	40 µL

Incubate: **30 min at 20 °C** followed by **30 min at 65 °C**, cool to 4°C.

Proceed directly to adapter ligation.

IMPORTANT: add 2 µL P5 modified adapter (mP5 adapter) to the DNA sample and mix well before adding ligation master mix! Adapter working concentration is 10- 20 µM.

Libraries with <2000/1500 ng: It is often recommended to aim for a 1:20 insert:adapter ratio calculated in moles. Thus, less adapter can be added for samples with low amount of starting material.

Library Building – Adapter Ligation (with mP5 adapter only)

2) Add the following components to the 40 µL end repaired DNA from above.

Ligase master mix

Reagent	Stock conc.	Final conc.	Total U	V/R µL	x
T4 DNA ligase buffer (NEB)	10x	1x		1	
PEG-4000	50%	6.25%		6	
T4 DNA ligase (NEB 400 U/µL)	400 U/µL	8	400	1	
Total:				**8**	

Reaction size:	50 µL

Incubate: 30 min at 20 °C followed by **10 min at 65 °C,** cool to 4°C.

3) Purify using 1.6x of Speedbeads. Elute in 30µL depending on initial input and expected outcome.

Real-time PCR

qPCR mix

Reagent	**# of reactions x1**	**# of reactions x N**	
	Volume (µl)		**Final Conc. X1**
H_2O	6.2		
2X KAPA HiFi HotStart Uracil+ ReadyMix	10		1X
BSA	0.8		1%
SYBR Green/ROXY	0.8		
Forward primer (IS4 10µM)	0.6		0.3µM
Reverse Alu primer (10µM)	0.6		0.3µM
Total master vol.	19 µl		
DNA template		1 µl	
Final vol.		**20 µl**	

qPCR thermal profile

Initial denaturation		95ºC	1 cycle	03:00
Incubation	Denature	98ºC	*	00:20
	Anneal	65ºC	*35 cycles	00:15
	Extend	72ºC	*	00:30
Final extension		72ºC	1 cycle	03:00

NOTE: The next steps are the TE-target enrichment PCR and the indexing PCR. They can be performed in one or two days, depending on the amount of

samples that you are working with, the purification set up (strips vs single tubes), or your lab set up.

TE-target PCR (PCR 1)

- **KAPA HiFi**

1) Prepare the TE PCR.

KAPA HiFi DNA Polymerase master mix

Reagent	**Reaction X1**	**Reactions X N**	
	Volume (µl)		**Final Conc. X1**
2X KAPA HiFi HotStart Uracil+ ReadyMix	10		1X
Forward primer (IS4) (10µM)	0.6		0.3µM
Reverse TE target primer (10µM)	0.6		0.3µM
BSA (30%)	1.33		2%
Molecular grade water	2.47		
Total	**15**		

IMPORTANT: The reverse TE primer is specific for the target species/ TE.

2) Add 5 µL of library to 20 µL of PCR reaction mix.
3) Amplify each of the master mixes using the following cycling conditions.

KAPA HiFi DNA polymerase thermal profile

95 °C – 3 minutes

98 °C – 20 seconds
67 °C* – 15 seconds x 22- 30 cycles (according to qPCR Ct/ Cp values)
*This temperature is TE- target primer dependent. Usually around 60- 65°C works well

72 °C – 30 seconds

72 °C – 3 min

10 °C – Hold

3) Purify the amplified library using 1.8x of SPEEDbeads as previously described.
4) Elute the sample in 33 µL EB (collect 30 µl).

Index of the Library (PCR 2)

- **KAPA HiFi**

1) Prepare the indexing PCR.

KAPA HiFi DNA Polymerase master mix

Reagent	**Reaction X1**	**Reactions X N**	
	Volume (µl)		**Final Conc. X1**
2X KAPA HiFi HotStart Uracil+ ReadyMix	10		1X
BSA (30%)	1.33		2%
Molecular grade water	1.67		
Total	**13**		

INDEXING:

For high multiplexing and dual indexing, use a system of P5* P7 indexing primers (10µM). Use 1 µl of each index.

2) Add 5 µL of purified material to the mastermix and amplify for 5- 10 cycles using the following conditions (cycle numbers can depend on the success of the first PCR).

KAPA HiFi DNA polymerase thermal profile

95 °C – 3 minutes

98 °C – 20 seconds

65 °C* – 15 seconds x 5- 10 Cycles

72 °C – 30 seconds

72 °C – 3 min

10 °C – Hold

3) Purify the amplified library using 1.8x of SPEEDbeads as previously described.
4) Elute the sample in 33 µL EB (collect 30 µl).
5) Measure on Qubit and Bioanalyzer/TS.

Appendix:

Oligos

Name 5'-3'

IS1_adapter P5 A*C*A*C*TCTTTCCCTACACGACGCTCTTCCG*A*T*C*T (Meyer and Kircher 2010)

mIS3_adapter P5+P7 (no 5' PHO)A*G*A*T*CGGAA*G*A*G*C****C3***

IS7 ACACTCTTTCCCTACACGAC (Carøe et al. 2018)

IS4 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT (Meyer and Kircher 2010)

P5_Index

AATGATACGGCGACCACCGAGATCTACACnnnnnnACACTCTTTCCCTACACGAC GCTCT (Carøe et al. 2018)

P7_Index

CAAGCAGAAGACGGCATACGAGATnnnnnnGTGACTGGAGTTCAGACGTGT (Meyer and Kircher 2010)

TE-target primer (P7+TE-target)

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT+TEsequence (18 & 24 bp)

TE-sequence:

Alu_v1 ATTACAGGCGTGAGCCACCGCGCC

Alu_v2 TGAGCCACCGCGCCCGGC

- * = phosphothioate linkage,
- n = index base
- [C3spacer] = C3 blocks extension

Preparation of the mP5 adapter

IS1_adapter P5 and mIS3_adapter P5+P7.

Synthesize to minimum 0.2 µmol scale, HPLC purify. Ship dry. Dissolve in 10 mM Tris-HCl to 500 µM.

Hybridization mix for modified P5 (200 µM)	
IS1 adapter P5.F (500 µM)	40 µl
mIS3 P5+P7 (500 µM)	40 µl
Oligo hybridization buffer (10x)	10 µl
H_2O	10 µl
Total volume	**100 µl**

Oligo hybridization buffer (10X) is prepared as in (Meyer and Kircher 2010) **(for adapter preparation):**

- 500 mM NaCl
- 10 mM Tris-HCl, pH 8.0
- 1 mM EDTA, pH 8.0

Mix and incubate the tube at 95 °C for 10 sec followed by a ramp from 95 °C to 12 °C at a rate of 0,1 °C/sec. The final adapter mix has a concentration of 200 µM of each adapter. Dilute with EBT to desired working solution e.g. 25 µM.

TE buffer (for SPRIbead preparation)

EDTA (0.5 M, pH 8.0) 0.2 mL (final conc. 1 mM)

Tris-HCl (1 M, pH 8.0) 1 mL (final conc. 10 mM)

H_2O to 100 mL

EBT buffer (for elution, dilution etc.)

1. 10 mM Tris-Cl, pH 8.0 or 8.5 (identical to QIAGEN's EB buffer)
2. 0.05% Tween 20

Reaction enhancer (for End-Repair reaction)

1. 0.25 g PEG (25% final concentration)
2. 100 µL BSA (20 mg/mL) (2 mg/mL final concentration)

3. 80 µL NaCl (5M) (400 mM final concentration) H_2O up to 1 mL

Can be frozen, used directly in reaction or mixed with dNTP and T4 DNA Ligase buffer to form a readymade mastermix.

SPRI bead preparation (SpeedBeads)

1. Mix Sera-mag SpeedBeads and transfer 1mL to a 1.5 mL Eppendorf tube.
2. Place the Eppendorf on a magnet rack and wait 30 seconds for the liquid to be clear.
3. Discard the supernatant
4. Remove the Tube from the magnet rack and add 1 mL TE buffer and bring the beads into solution by flicking the tube.
5. Place on magnetrack again, wait for the liquid to be clear (app. 30 seconds), discard the supernatant
6. Repeat the wash (step 4- 5)
7. Resuspend the beads in 1 mL TE buffer and place in NON-magnetic rack.
8. Add 9 grams of PEG-8000 to a 50 mL tube.
9. Add 10 mL (5M) NaCL (or 2.92g) to the 50 mL tube.
10. Add 500 µL 1M Tris-HCL to the 50 mL tube.
11. Add 100 µL 0.5 M EDTA to the 50 mL tube.
12. Add molecular grade water to 49 mL.
13. Mix solution until all PEG has been dissolved. (app. 5 minutes). Shake well.
14. Add 27.5 µLTween 20 to the PEG solution (pipette slowly).
15. Add the 1 mL prepared SpeedBead/ TE solution to the PEG solution to reach 50 mL final Bead solution. Mix gently to obtain a uniform color.
16. Aliquot in smaller tubes for ease of use and store dark at 4°C.

Clean up the library with SPRI bead purification according to (Rohland and Reich 2012)**.** SPRI bead purification has been successfully tested by adding 1.6x of SPRI bead solution to the final library, incubating for 5 minutes before washing beads twice with 80% ethanol (without removing plate from magnet), dry them (do not overdry!) and elute in 33 µL EB, by incubation for 10 minutes at 37°C, before collecting 30 µl of supernatant.

Reagents

Enzymes:

- T4 Polynucleotide Kinase; Sigma-Aldrich cat#M0201S; size: 500 units; 10,000 units/ml; Price $56.00
- T4 DNA Polymerase; Sigma-Aldrich cat#M0203S; size 150 units; 3,000 units/ml; price $63.00
- T4 DNA Ligase; Sigma-Aldrich Cat#M0202S; size 20,000 units; 400,000 units/ml; price $64.00
- KAPA Taq HotStart with dNTPs; Sigma-Aldrich Cat#KK1511; size 500 U; 5,000 units/ml; price $242.00

Other reagents:

- SpeedBeads™ magnetic carboxylate modified particles, Sigma-Aldrich, Cat#: GE45152105050250
- PEG-8000, Sigma- Aldrich, Cat#: 89510-250G-F (for SPRI bead solution)
- PEG-4000, Sigma- Aldrich Cat#: 95904-250G-F (for library build)
- 0.5M EDTA, pH8.0, e.g. Sigma- Aldrich
- 1.0M Tris, pH8.0, e.g. Sigma- Aldrich
- Tween 20– 50% e.g. Sigma- Aldrich
- 5M NaCl solution, e.g. Sigma- Aldrich
- EB elution buffer, Qiagen Cat#: 19086
- Tris-EDTA buffer solution,e.g. Sigma-Aldrich Cat#: T9285
- T4 DNA ligase buffer (NEB) (included with the ligase (Sigma-Aldrich Cat#M0202S))
- PCR buffer II (included with the DNA polymerase (Thermo Fisher Scientific Cat#N8080241))
- $MgCl_2$ Solution (included with the DNA polymerase (Thermo Fisher Scientific Cat#N8080241))
- BSA (20mg/mL), e.g. Sigma- Aldrich
- dNTPs solution, e.g. Thermo Fisher Scientific
- Molecular grade water, e.g. Sigma- Aldrich

Plastic ware:

- Filtertips, 0.1-1 mL, 20-200 µL, 2-20 µL and 1-10 µL

- Eppendorf 0.5 ml Lobind tubes (for doing few samples, convenient for ancient DNA)
- Eppendorf Lobind 96 well plate (if doing many samples)
- PCR strips (for doing few or many samples)
- 50 mL Falcon Tubes (to prepare speedbeads)
- Zip lock bags or boxes for freezer storage, and pen for labeling.
- Waste bags and small bin for tips etc.

Other lab equipment:

Latex or nitril gloves (and suit, mouth cover etc. for ancient DNA)
Small table centrifuge for 1.5 mL eppendorf tubes
Single channel pipettes: (e.g. Thermo-Fisher Scientific Finnpippete system F1)
Multi Channel pipettes: E.g. e.g. Thermo-Fisher Scientific Finnpippete system F1, 30-300 µL, (for many samples and plate or strip setup)
Racks for 0.5 Eppendorf tubes.
Magnetic Rack for 1,5 mL Eppendorf tubes (for speedbead preparation)
Magnetic plate for 96 well plates (for speedbead cleanup)
Cooling block for 96 well plates or PCR strips (when setup up reactions)
Ice tray with ice (for reagents or samples when setting up reactions)
Cooling ThermoMixer MKR13 (HLC/ Ditabis) (for adapter preparation)
Thermocycler E.g. Applied Biosystems 2720 Thermal Cycler (for incubation)
Equipment for sanitation, such as bleach (hypochlorite), UV-Crosslinker and 70% ethanol and paper cloth (for sanitation and cleaning).

References

- Carøe, Christian, Shyam Gopalakrishnan, Lasse Vinner, Sarah S. T. Mak, Mikkel Holger S. Sinding, José A. Samaniego, Nathan Wales, Thomas Sicheritz-Pontén, and M. Thomas P. Gilbert. 2018. “Single-Tube Library Preparation for Degraded DNA.” *Methods in Ecology and Evolution / British Ecological Society 9(2),* 410–19.
- Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols, 2010(6)*, pdb-prot5448.
- Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research, 22,* 939–946.

Appendix S3- TapeStation results

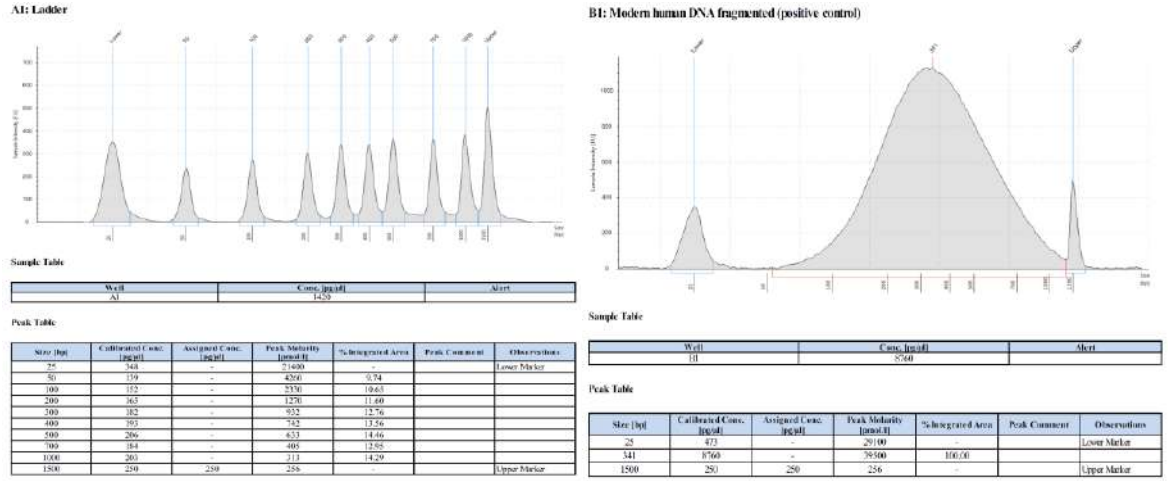
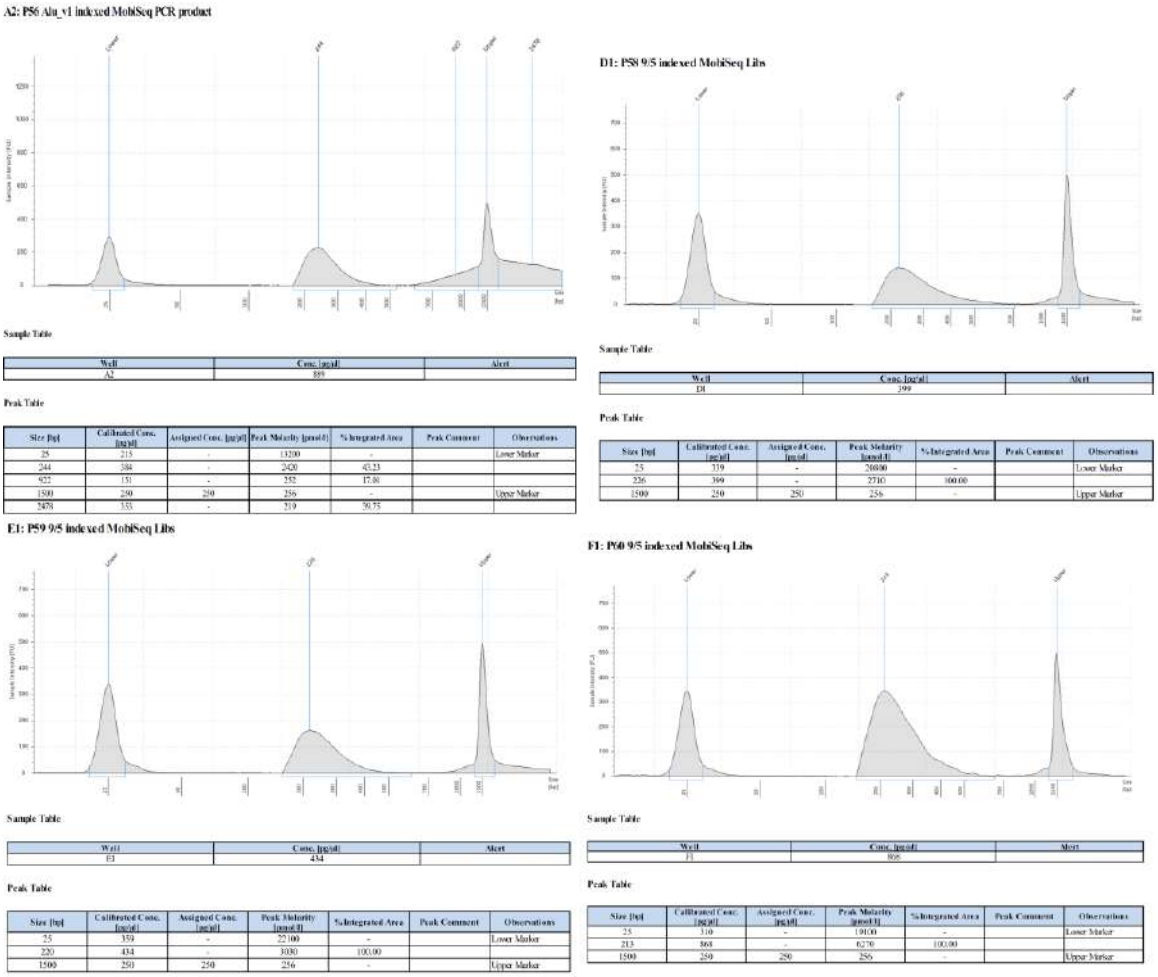
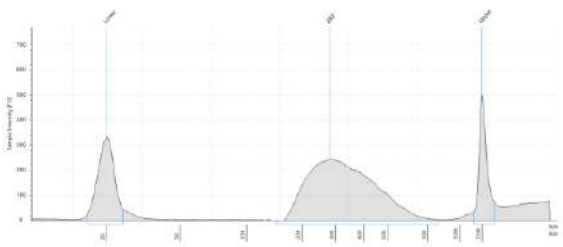


Figure S5. TapeStation results for positive control before MobiSeq and the profile of the High sensitivity DNA ladder.



G1: P61 9/5 indexed MobiSeq Libs



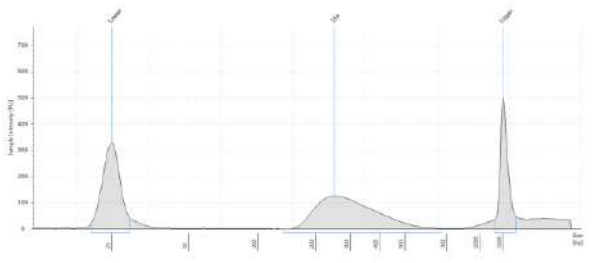
Sample Table

Well	Conc. [ng/ul]	Alert
G1	1060	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	351	-	21600	-		Lower Marker
282	1060	-	5750	100.00		
1500	250	250	256	-		Upper Marker

H1: P62 9/5 indexed MobiSeq Libs



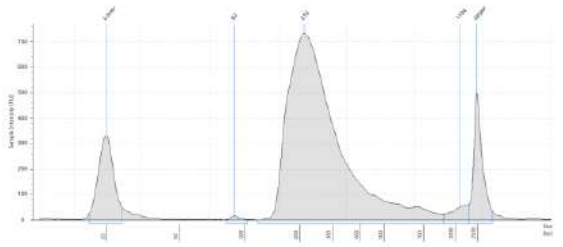
Sample Table

Well	Conc. [ng/ul]	Alert
H1	472	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	374	-	23000	-		Lower Marker
254	472	-	2860	100.00		
1500	250	250	256	-		Upper Marker

A2: P63 alu-v1 9/5 indexed MobiSeq Libs



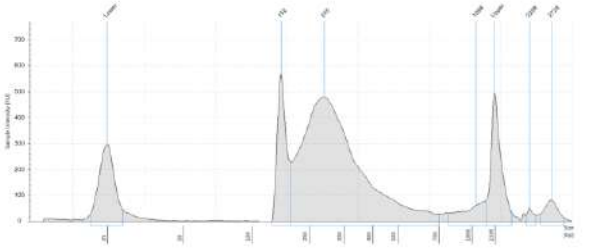
Sample Table

Well	Conc. [ng/ul]	Alert
A2	2180	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	297	-	18500	-		Lower Marker
92	7.33	-	122	0.34		
215	2120	-	15200	97.24		
1104	52.8	-	73.5	2.42		
1500	250	250	256	-		Upper Marker

B2: P63 alu-v2 9/5 indexed MobiSeq Libs



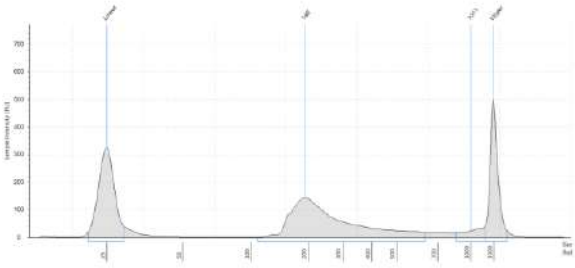
Sample Table

Well	Conc. [ng/ul]	Alert
B2	1750	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	232	-	14300	-		Lower Marker
152	264	-	2670	15.03		
245	1350	-	8440	76.70		
1008	78.6	-	110	4.48		
1500	250	250	256	-		Upper Marker
2258	14.2	-	9.69	0.81		
2729	52.2	-	29.4	2.98		

C2: P64 9/5 indexed MobiSeq Libs



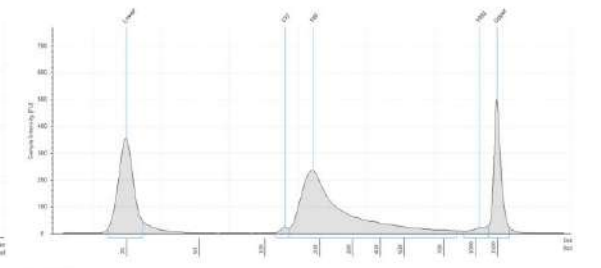
Sample Table

Well	Conc. [ng/ul]	Alert
C2	524	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	334	-	20600	-		Lower Marker
195	484	-	3820	92.42		
1011	39.8	-	60.5	7.58		
1500	250	250	256	-		Upper Marker

D2: P65 9/5 indexed MobiSeq Libs



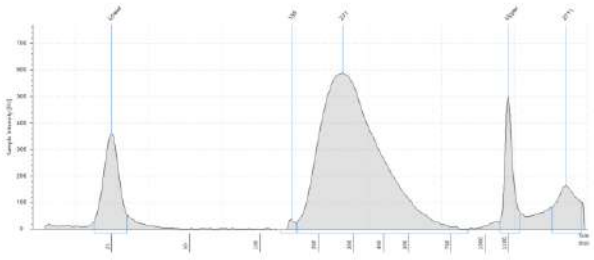
Sample Table

Well	Conc. [ng/ul]	Alert
D2	707	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	374	-	23000	-		Lower Marker
137	12.5	-	140	1.76		
189	669	-	5450	94.63		
1002	25.5	-	35.9	3.60		
1500	250	250	256	-		Upper Marker

E2: P66 9/5 indexed MobiSeq Libs



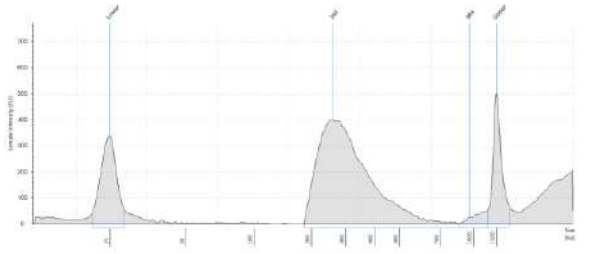
Sample Table

Well	Conc. [pg/ul]	Alert
E2	2440	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	360	-	22500	-		Lower Marker
155	12.7	-	127	0.52		
271	2220	-	12600	91.01		
1500	250	250	256	-		Upper Marker
2711	207	-	117	8.47		

F2: P67 9/5 indexed MobiSeq Libs



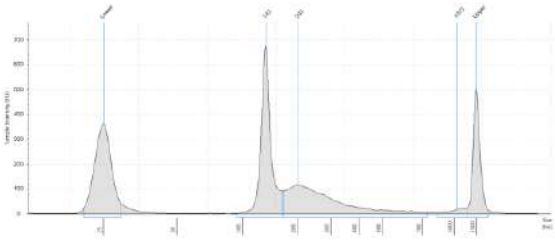
Sample Table

Well	Conc. [pg/ul]	Alert
F2	1190	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	315	-	19400	-		Lower Marker
264	1150	-	6720	96.59		
964	40.7	-	65.0	3.41		
1500	250	250	256	-		Upper Marker

G2: P68 9/5 indexed MobiSeq Libs



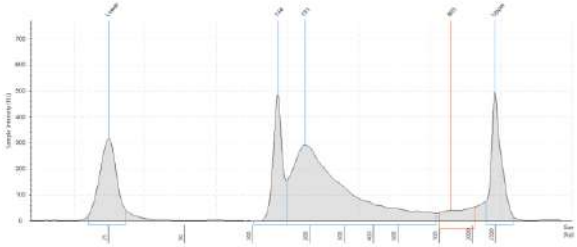
Sample Table

Well	Conc. [pg/ul]	Alert
G2	885	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	433	-	26600	-		Lower Marker
143	463	-	4990	52.40		
203	395	-	3990	44.68		
1072	25.8	-	37.0	2.92		
1500	250	250	256	-		Upper Marker

H2: P69 9/5 indexed MobiSeq Libs



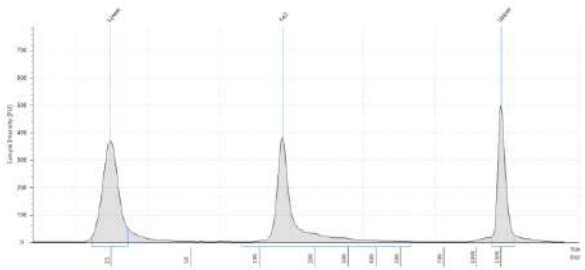
Sample Table

Well	Conc. [pg/ul]	Alert
H2	1050	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	272	-	16700			Lower Marker
144	226	-	2350	26.91		
191	288	-	6180	72.07		
895	6.14	-	123	1.12		
1500	250	250	256	-		Upper Marker

B1: P70 9/5 indexed MobiSeq Libs



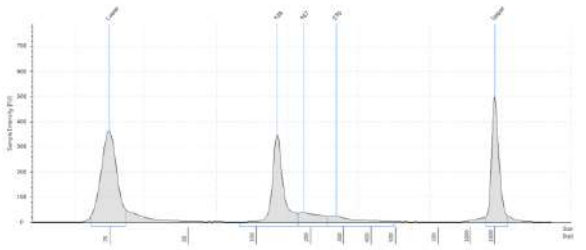
Sample Table

Well	Conc. [pg/ul]	Alert
B1	370	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	380	-	23400	-		Lower Marker
142	370	-	4000	100.00		
1500	250	250	256	-		Upper Marker

D1: P72 9/5 indexed MobiSeq Libs



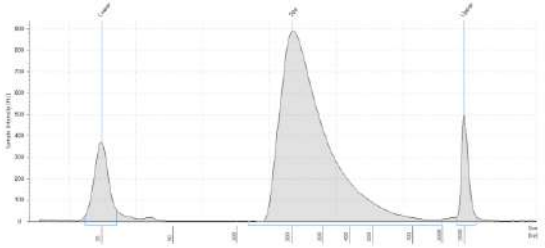
Sample Table

Well	Conc. [pg/ul]	Alert
D1	344	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	382	-	23500			Lower Marker
130	249	-	2750	72.23		
187	54.5	-	447	15.82		
278	41.1	-	228	11.94		
1500	250	250	256	-		Upper Marker

D1: P73 aluv1 index PCR product



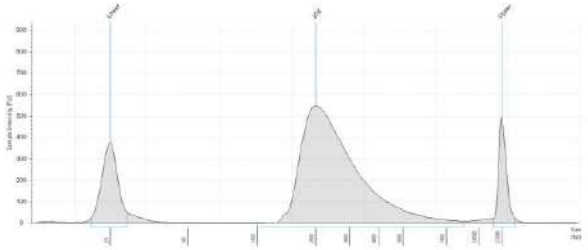
Sample Table

Well	Conc. [pg/µl]	Alert
D1	2860	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	385	-	23700	-		Lower Marker
204	2860	-	21600	100.00		
1500	250	250	256	-		Upper Marker

E1: P73 aluv2 index PCR product



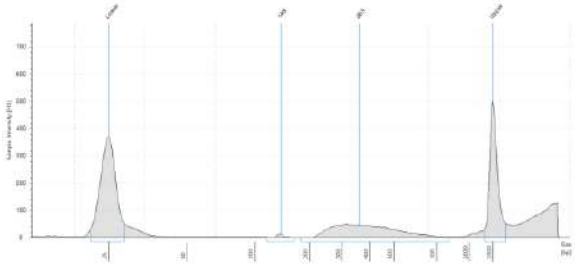
Sample Table

Well	Conc. [pg/µl]	Alert
E1	1910	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	391	-	24000	-		Lower Marker
202	1910	-	14500	100.00		
1500	250	250	256	-		Upper Marker

E1: P74 9/5 indexed MobiSeq Libs



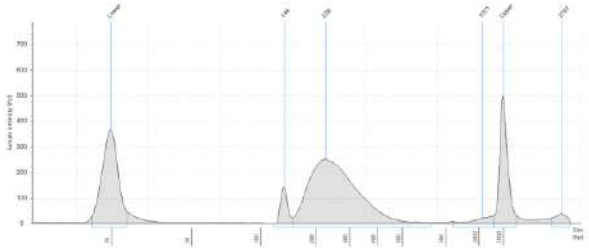
Sample Table

Well	Conc. [pg/µl]	Alert
E1	201	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	366	-	22500	-		Lower Marker
148	4.22	-	43.8	2.10		
363	196	-	833	97.90		
1500	250	250	256	-		Upper Marker

F1: P75 9/5 indexed MobiSeq Libs



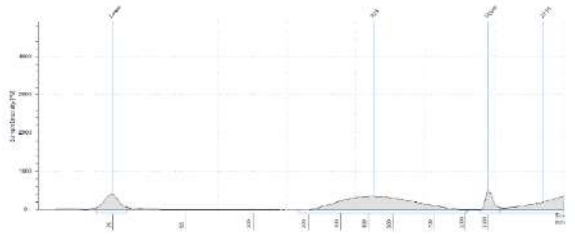
Sample Table

Well	Conc. [pg/µl]	Alert
F1	985	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	270	-	22700	-		Lower Marker
144	63.8	-	684	7.21		
229	760	-	5110	85.90		
1071	30.5	-	43.7	3.44		
1500	250	250	256	-		Upper Marker
2707	30.5	-	17.3	3.44		

D2: Positive Control (17/4/19) Alu_v1 Indexed MobiSeq PCR products



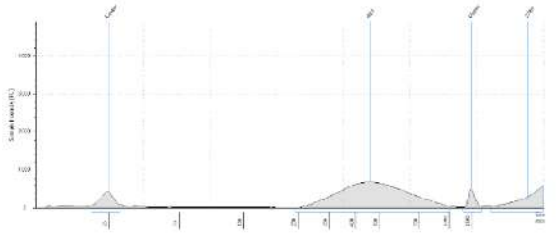
Sample Table

Well	Conc. [pg/µl]	Alert
D2	2320	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	406	-	21600	-		Lower Marker
423	1350	-	6303	75.43		
1500	250	250	256	-		Upper Marker
2718	566	-	322	24.57		

E2: Positive Control (17/4/19) Alu_v2 Indexed MobiSeq PCR products



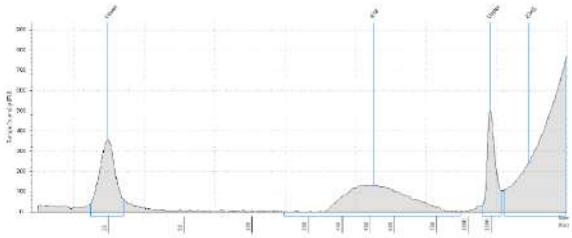
Sample Table

Well	Conc. [pg/µl]	Alert
E2	4430	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	448	-	23600	-		Lower Marker
461	3680	-	12300	82.69		
1500	250	250	256	-		Upper Marker
2707	770	-	475	17.31		

F1: Positive control alu1 index PCR product 5/4/2019



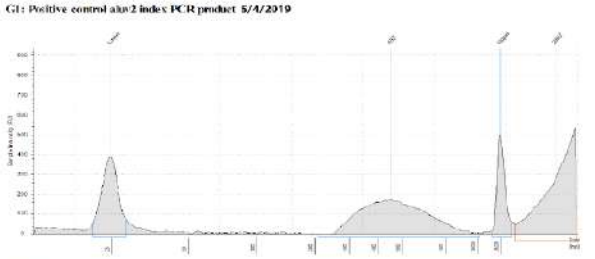
Sample Table

Well	Conc. [pg/ul]	Alert
F1	1630	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	346	-	21300	-		Lower Marker
416	597	-	3860	30.42		
1500	256	250	256	-		Upper Marker
2348	1150	-	764	69.58		

G1: Positive control alu2 index PCR product 5/4/2019



Sample Table

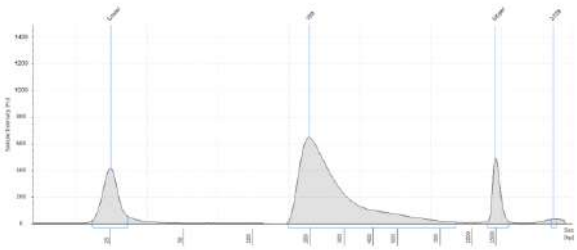
Well	Conc. [pg/ul]	Alert
G1	1520	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
35	405	-	24900	-		Lower Marker
457	759	-	2401	46.48		
1501	750	750	216	-		Upper Marker
3607	705	-	449	53.52		

Figure S6. TapeStation results for MobiSeq libraries using both Alu TE-target primers.

B1: PS6 BST PCR product



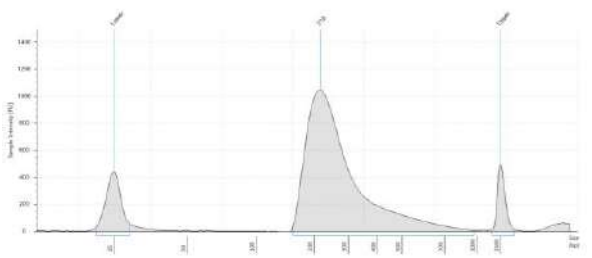
Sample Table

Well	Conc. [pg/ul]	Alert
B1	1770	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	427	-	26300	-		Lower Marker
199	1760	-	13600	99.40		
1500	250	250	256	-		Upper Marker
2728	10.7	-	6.91	0.60		

C1: PS7 BST PCR product



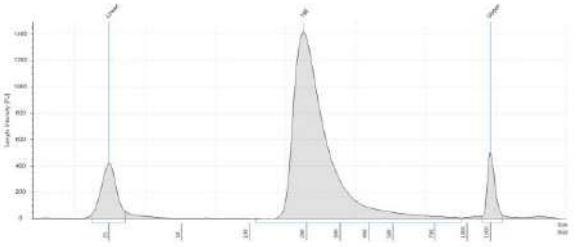
Sample Table

Well	Conc. [pg/ul]	Alert
C1	3070	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	428	-	26400	-		Lower Marker
218	3070	-	21600	100.00		
1500	250	250	256	-		Upper Marker

D1: PS8 BST PCR product



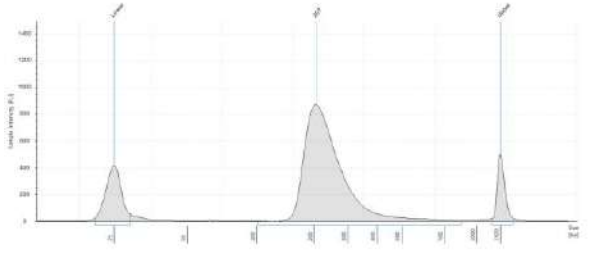
Sample Table

Well	Conc. [pg/ul]	Alert
D1	3080	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	419	-	25800	-		Lower Marker
195	3080	-	23700	100.00		
1500	250	250	256	-		Upper Marker

E1: PS9 BST PCR product



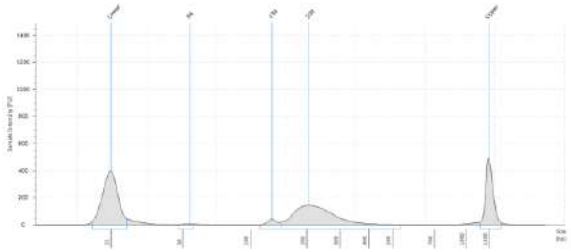
Sample Table

Well	Conc. [pg/ul]	Alert
E1	2050	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	427	-	26300	-		Lower Marker
207	2050	-	15200	100.00		
1500	250	250	256	-		Upper Marker

F1: P60 BST PCR product



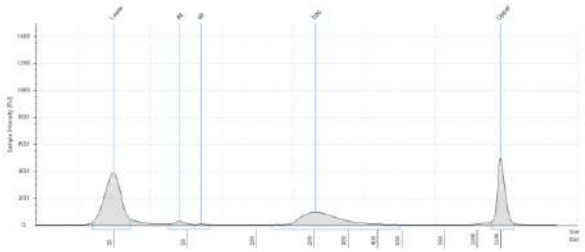
Sample Table

Well	Conc. [pg/µl]	Alert
F1	425	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	421	-	25900	-		Lower Marker
56	5.82	-	161	1.37		
139	28.0	-	311	6.60		
209	391	-	2890	92.03		
1500	250	250	256	-		Upper Marker

G1: P61 BST PCR product



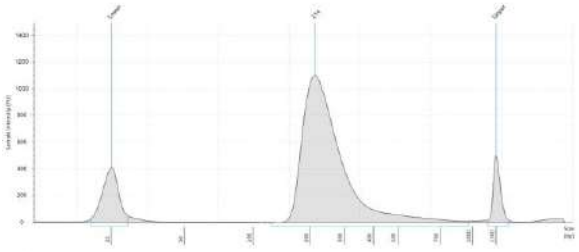
Sample Table

Well	Conc. [pg/µl]	Alert
G1	265	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	430	-	26800	-		Lower Marker
48	24.5	-	791	8.58		
60	5.16	-	132	1.81		
206	256	-	1910	89.61		
1500	250	250	256	-		Upper Marker

H1: P62 BST PCR product



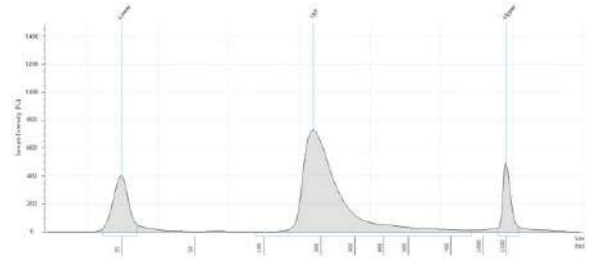
Sample Table

Well	Conc. [pg/µl]	Alert
H1	2880	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	463	-	28500	-		Lower Marker
214	2880	-	26800	100.00		
1500	250	250	256	-		Upper Marker

A2: P63 BST PCR product



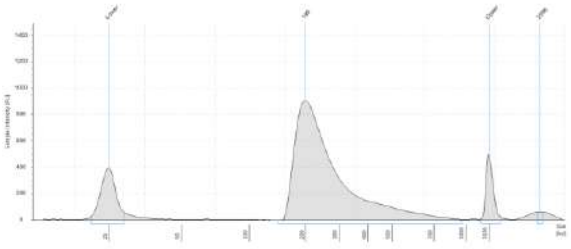
Sample Table

Well	Conc. [pg/µl]	Alert
A2	1650	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	406	-	25900	-		Lower Marker
187	1650	-	13500	100.00		
1500	250	250	256	-		Upper Marker

B2: P64 BST PCR product



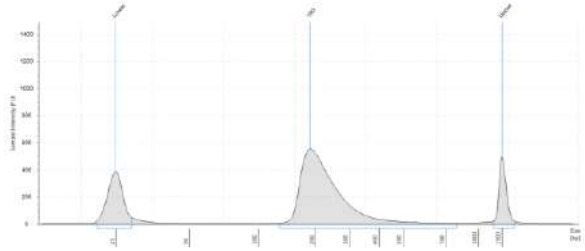
Sample Table

Well	Conc. [pg/µl]	Alert
B2	2680	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	420	-	25900	-		Lower Marker
199	2660	-	26500	99.25		
1500	250	250	256	-		Upper Marker
2596	20.6	-	12.2	0.77		

C2: P65 BST PCR product



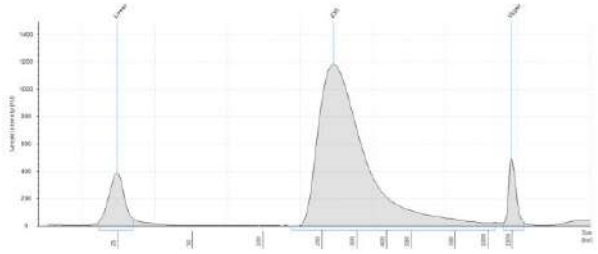
Sample Table

Well	Conc. [pg/µl]	Alert
C2	1370	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	411	-	25300	-		Lower Marker
190	1370	-	11100	100.00		
1500	250	250	256	-		Upper Marker

D2: P66 BST PCR product



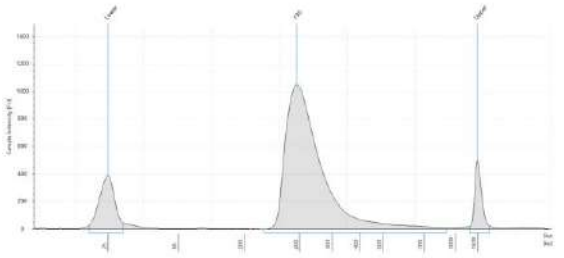
Sample Table

Well	Conc. [ng/ul]	Alert
D2	3680	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	400	-	25100	-		Lower Marker
235	3680	-	24100	100.00		
1500	250	250	256	-		Upper Marker

E2: P67 BST PCR product



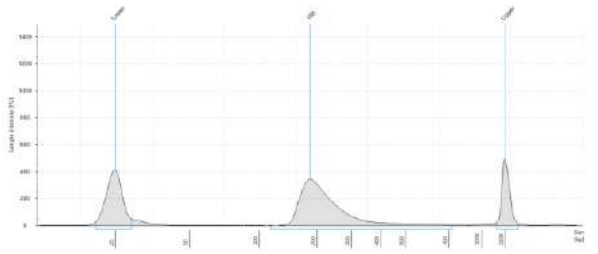
Sample Table

Well	Conc. [ng/ul]	Alert
E2	2680	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	434	-	26700			Lower Marker
193	2680	-	21100	100.00		
1500	250	250	256	-		Upper Marker

F2: P68 BST PCR product



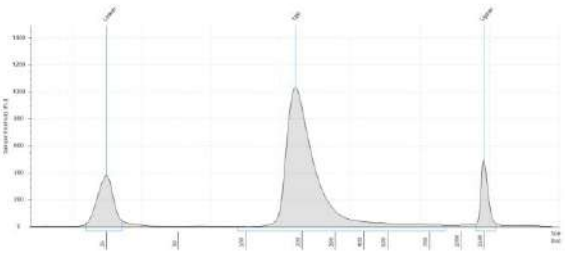
Sample Table

Well	Conc. [ng/ul]	Alert
F2	793	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	430	-	26500	-		Lower Marker
188	793	-	6480	100.00		
1500	250	250	256	-		Upper Marker

G2: P69 BST PCR product



Sample Table

Well	Conc. [ng/ul]	Alert
G2	2.60	

Peak Table

Size [bp]	Calibrated Conc. [pg/ul]	Assigned Conc. [pg/ul]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	434	-	26700			Lower Marker
189	2.60	-	16900	100.00		
1500	250	250	256	-		Upper Marker

H2: P70 BST PCR product

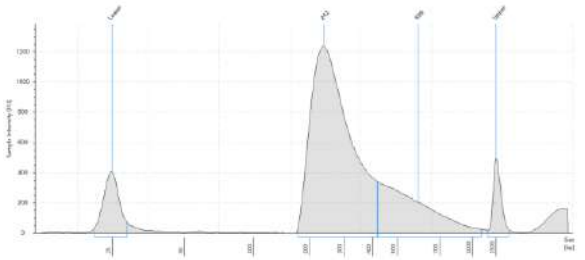
Sample Table

Well	Conc. [ng/ul]	Alert
H2	758	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	454	-	27900	-		Lower Marker
188	758	-	6210	100.00		
1500	250	250	256	-		Upper Marker

B1: P71 BST PCR product



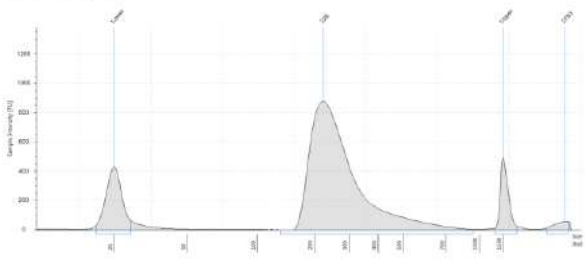
Sample Table

Well	Conc. [ng/ul]	Alert
B1	4.00	

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	413	-	25400	-		Lower Marker
242	3230	-	29400	76.56		
599	981	-	2520	23.44		
1500	250	250	256	-		Upper Marker

C1: P72 BST PCR product



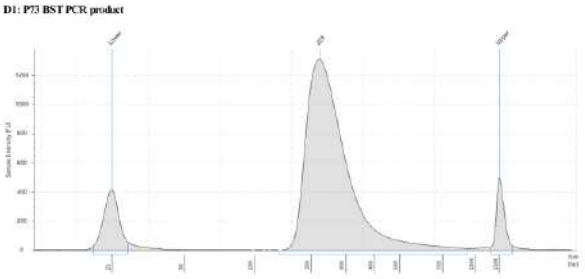
Sample Table

Well	Conc. [pg/µl]	Alert
C1	2490	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	437	-	26900	-		Lower Marker
226	2410	-	16600	98.07		
1500	250	250	256	-		Upper Marker
2793	47.9	-	28.4	1.93		

D1: P73 BST PCR product



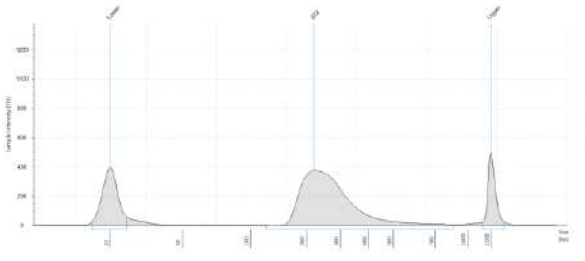
Sample Table

Well	Conc. [pg/µl]	Alert
D1	3340	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	405	-	25500			Lower Marker
224	3340	-	22900	100.00		
1500	250	250	256	-		Upper Marker

E1: P74 BST PCR product



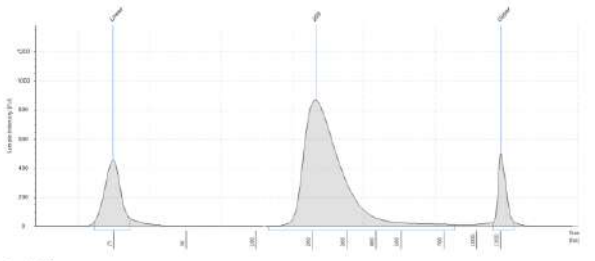
Sample Table

Well	Conc. [pg/µl]	Alert
E1	1220	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	423	-	26900	-		Lower Marker
225	1230	-	8450	100.00		
1500	250	250	256	-		Upper Marker

F1: P75 BST PCR product



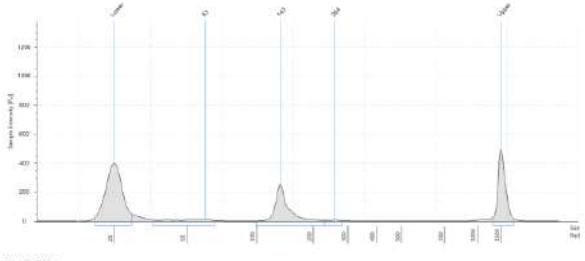
Sample Table

Well	Conc. [pg/µl]	Alert
F1	1040	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	458	-	28200			Lower Marker
209	1040	-	14700	100.00		
1500	250	250	256	-		Upper Marker

G1: Extraction blank BST PCR product



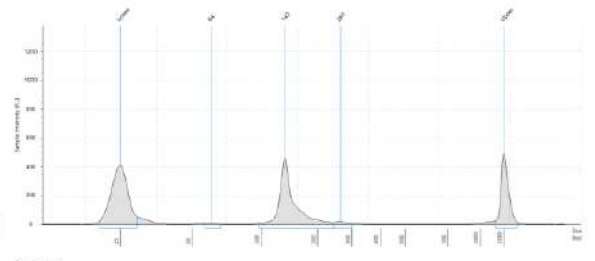
Sample Table

Well	Conc. [pg/µl]	Alert
G1	258	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	452	-	27000	-		Lower Marker
63	35.0	-	851	13.59		
143	215	-	2310	83.27		
264	14.9	-	47.2	3.14		
1500	250	250	256	-		Upper Marker

H1: BST Library blank PCR product



Sample Table

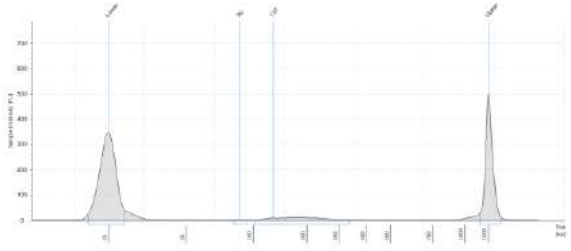
Well	Conc. [pg/µl]	Alert
H1	436	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	469	-	28900	-		Lower Marker
64	7.42	-	179	1.69		
142	416	-	4520	95.03		
267	14.4	-	92.9	5.28		
1500	250	250	256	-		Upper Marker

Figure S7. TapeStation results for shotgun libraries.

G1: P56 9/5 TE MobiSeq PCR product



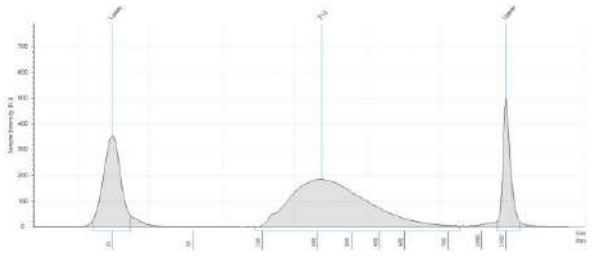
Sample Table

Well	Conc. [pg/µl]	Alert
G1	50.8	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	410	-	25200			Lower Marker
90	1.77	-	30.1	3.49		
137	49.0	-	551	96.51		
1500	250	250	256	-		Upper Marker

H1: P61 9/5 TE MobiSeq PCR product



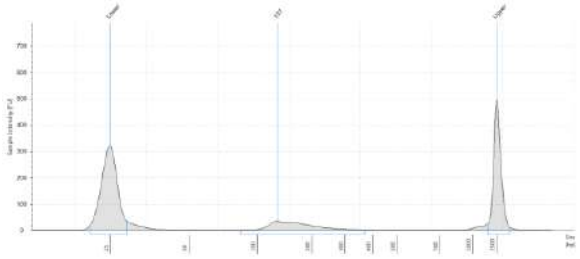
Sample Table

Well	Conc. [pg/µl]	Alert
H1	952	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	408	-	25100	-		Lower Marker
212	952	-	6010	100.00		
1500	250	250	256	-		Upper Marker

A2: P63 9/5 Alu-v1 TE MobiSeq PCR product



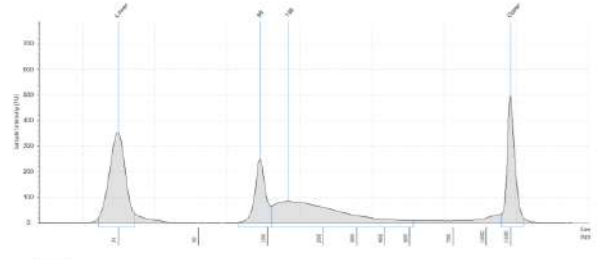
Sample Table

Well	Conc. [pg/µl]	Alert
A2	109	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	355	-	21900	-		Lower Marker
137	104	-	1180	100.00		
1500	250	250	256	-		Upper Marker

B2: P63 9/5 Alu-v2 TE MobiSeq PCR product



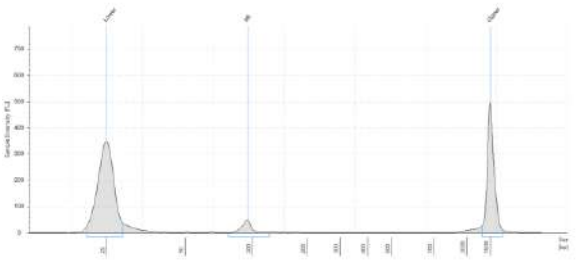
Sample Table

Well	Conc. [pg/µl]	Alert
B2	551	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	273	-	22900	-		Lower Marker
95	171	-	2770	32.19		
138	360	-	4010	67.81		
1500	250	250	256	-		Upper Marker

C2: P66 9/5 TE MobiSeq PCR product



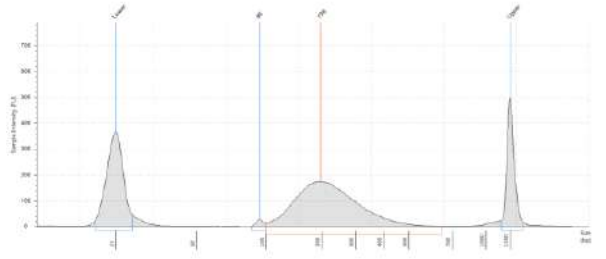
Sample Table

Well	Conc. [pg/µl]	Alert
C2	35.8	

Peak Table

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
25	421	-	25900	-		Lower Marker
96	35.8	-	572	100.00		
1500	250	250	256	-		Upper Marker

D2: P71 9/5 Alu-v1 TE MobiSeq PCR product



Sample Table

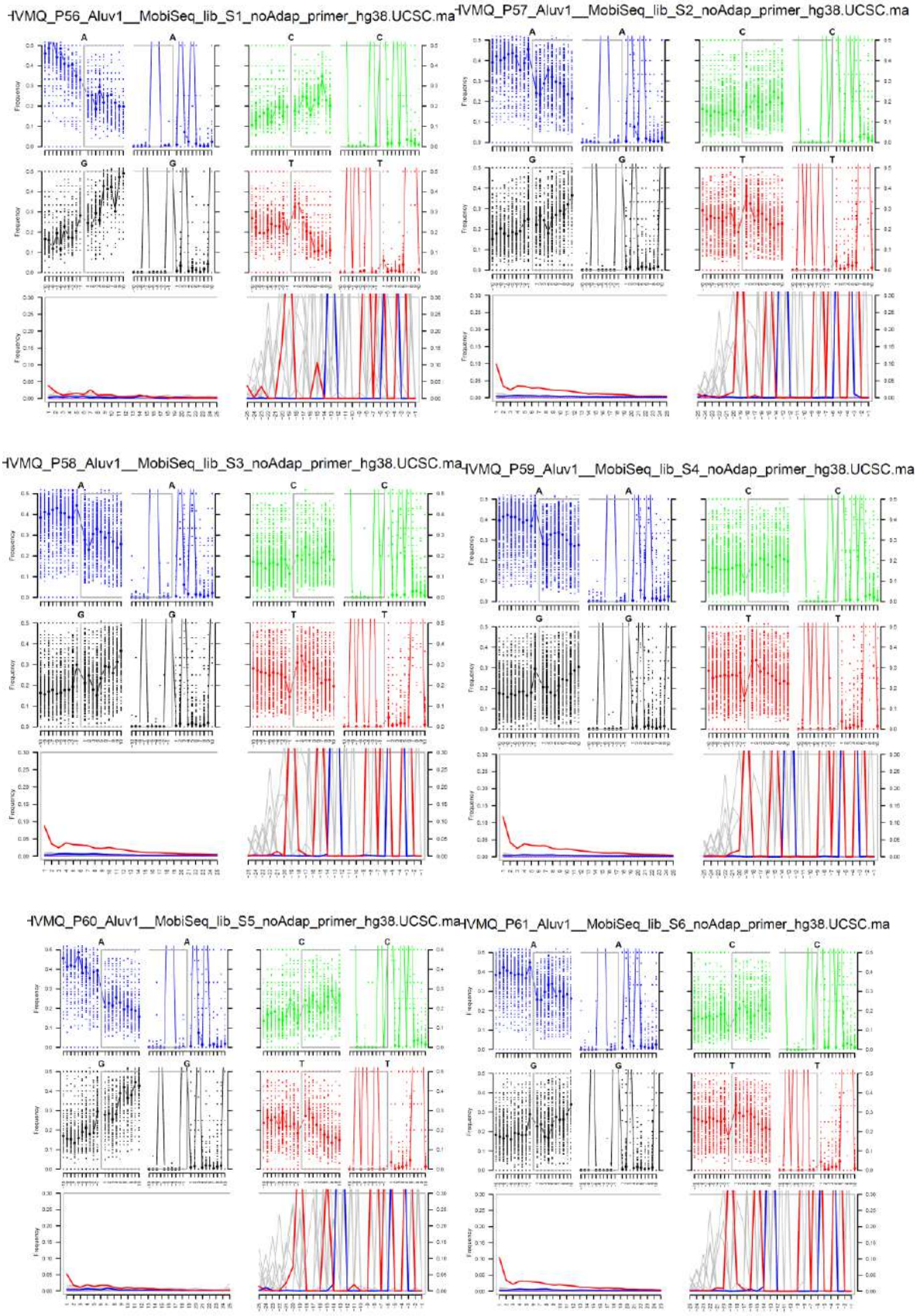
Well	Conc. [pg/µl]	Alert
D2	747	

Peak Table

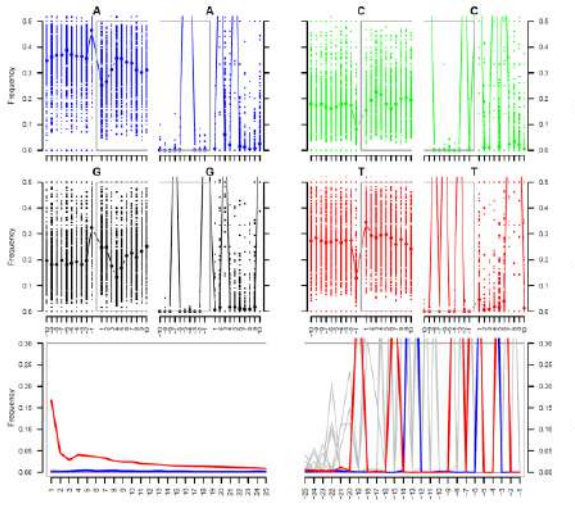
Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	%Integrated Area	Peak Comment	Observations
25	422	-	26000	-		Lower Marker
95	13.6	-	219	1.82		
198	733	-	5710	98.18		
1500	250	250	256	-		Upper Marker

Figure S8. TapeStation results for TE-PCR products using both Alu TE-target primers.

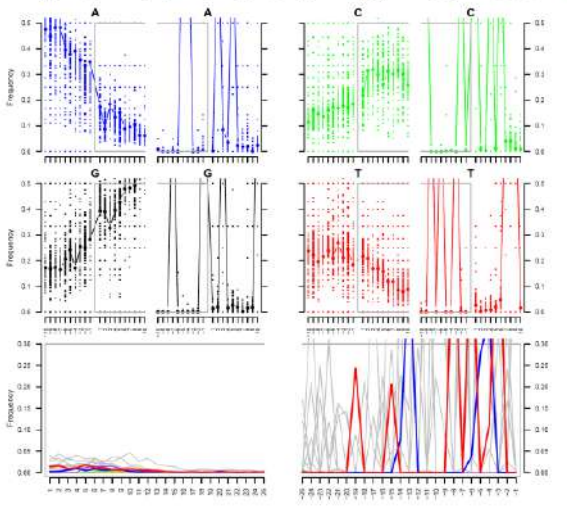
Appendix S4- MapDamage profiles



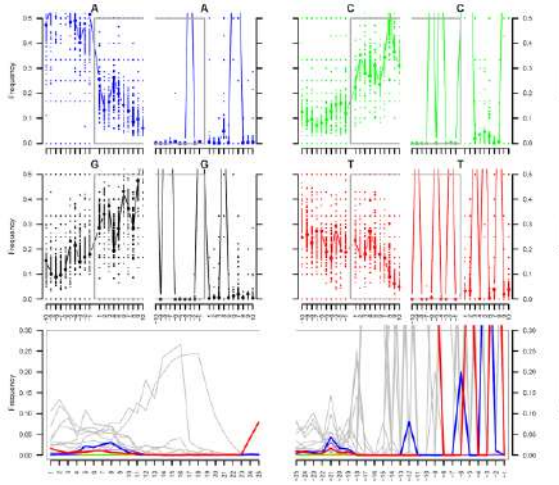
-IVMQ_P62_Aluv1__MobiSeq_lib_S7_noAdap_primer_hg38.UCSC.ma



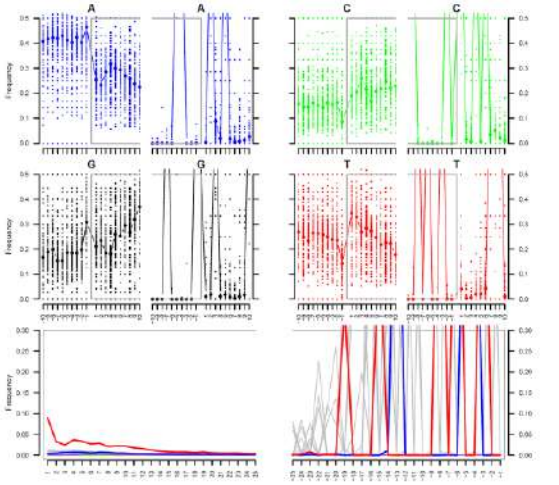
IVMQ_P63_Aluv1__MobiSeq_lib_S8_noAdap_primer_hg38.UCSC.ma



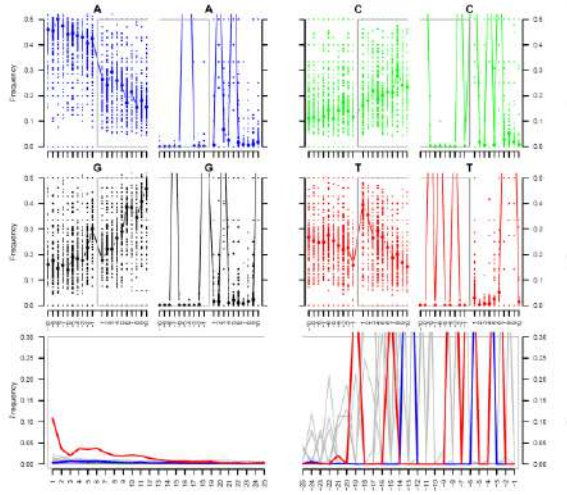
IVMQ_P63_Aluv2__MobiSeq_lib_S24_noAdap_primer_hg38.UCSC.ma



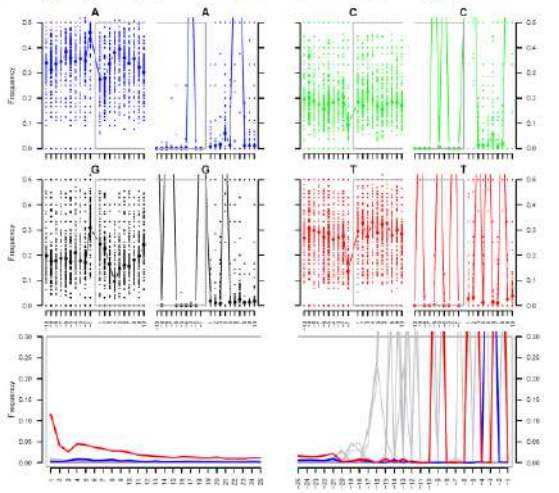
IVMQ_P64_Aluv1__MobiSeq_lib_S9_noAdap_primer_hg38.UCSC.ma



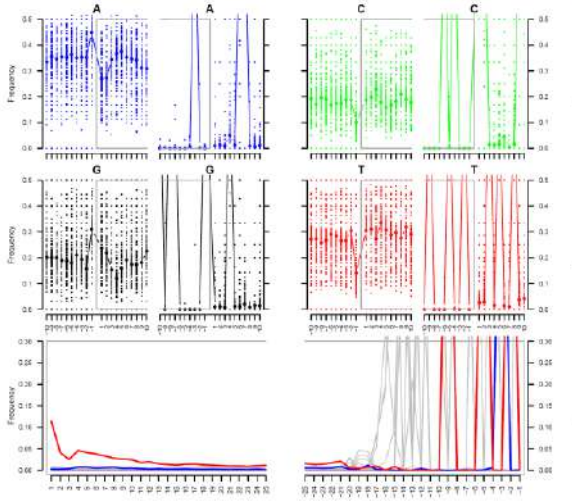
IVMQ_P65_Aluv1__MobiSeq_lib_S10_noAdap_primer_hg38.UCSC.ma



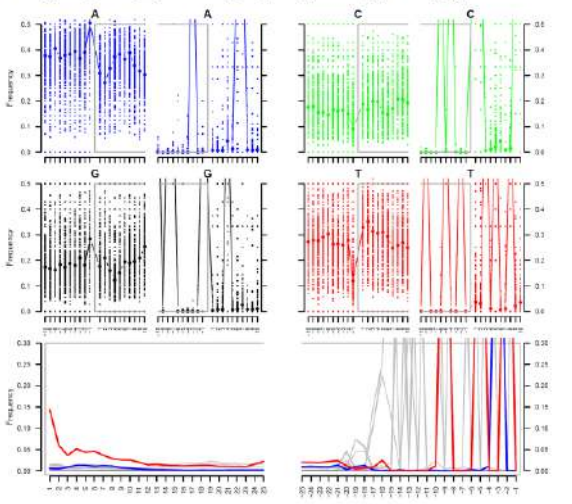
IVMQ_P66_Aluv2__MobiSeq_lib_S14_noAdap_primer_hg38.UCSC.ma



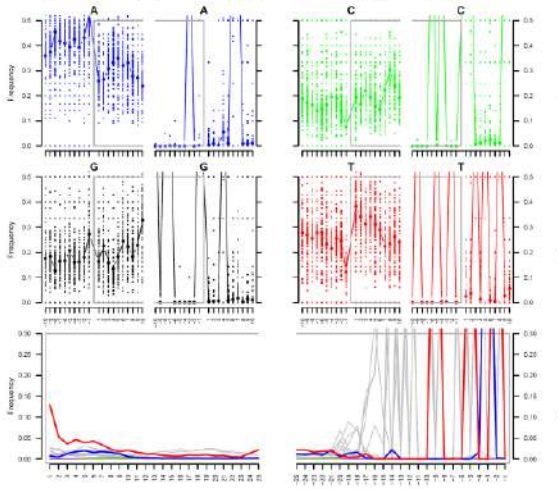
IVMQ_P67_Aluv2__MobiSeq_lib_S15_noAdap_primer_hg38.UCSC.ma



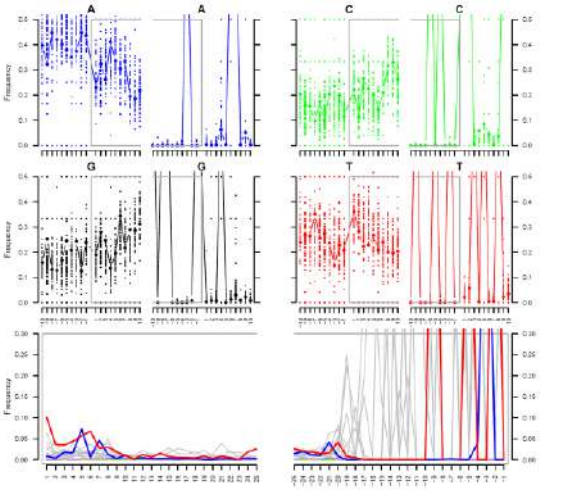
IVMQ_P68_Aluv2__MobiSeq_lib_S16_noAdap_primer_hg38.UCSC.ma



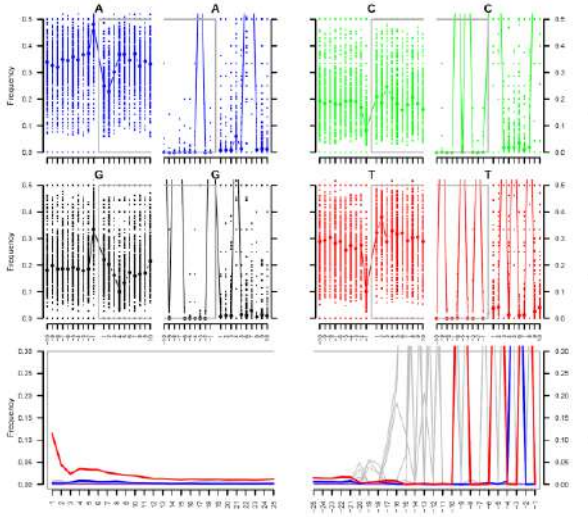
IVMQ_P69_Aluv2__MobiSeq_lib_S17_noAdap_primer_hg38.UCSC.ma



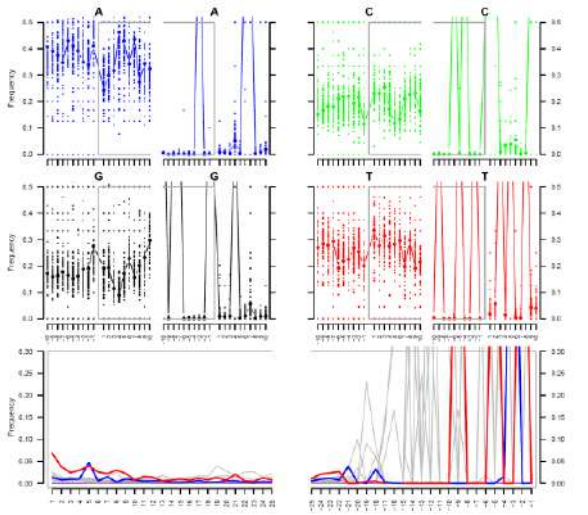
IVMQ_P70_Aluv2__MobiSeq_lib_S18_noAdap_primer_hg38.UCSC.ma



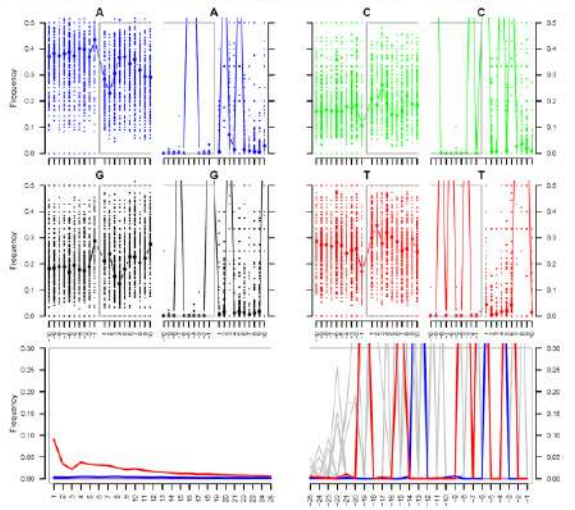
IVMQ_P71_Aluv2__MobiSeq_lib_S19_noAdap_primer_hg38.UCSC.ma



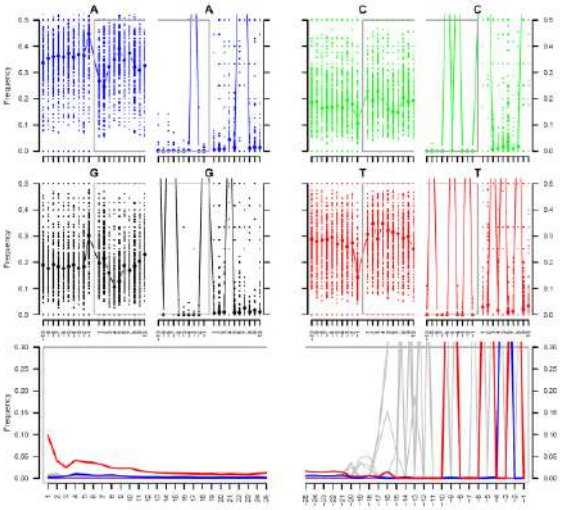
IVMQ_P72_Aluv2__MobiSeq_lib_S20_noAdap_primer_hg38.UCSC.ma



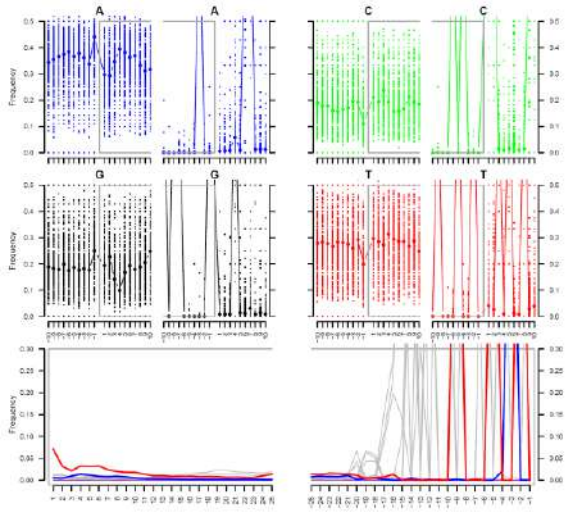
IVMQ_P73_Aluv1__MobiSeq_lib_S11_noAdap_primer_hg38.UCSC.ma



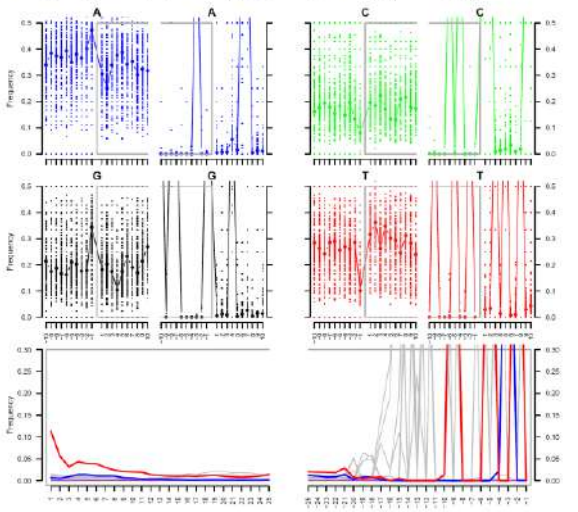
IVMQ_P73_Aluv2__MobiSeq_lib_S21_noAdap_primer_hg38.UCSC.ma



IVMQ_P74_Aluv2__MobiSeq_lib_S22_noAdap_primer_hg38.UCSC.ma



IVMQ_P75_Aluv2__MobiSeq_lib_S23_noAdap_primer_hg38.UCSC.ma



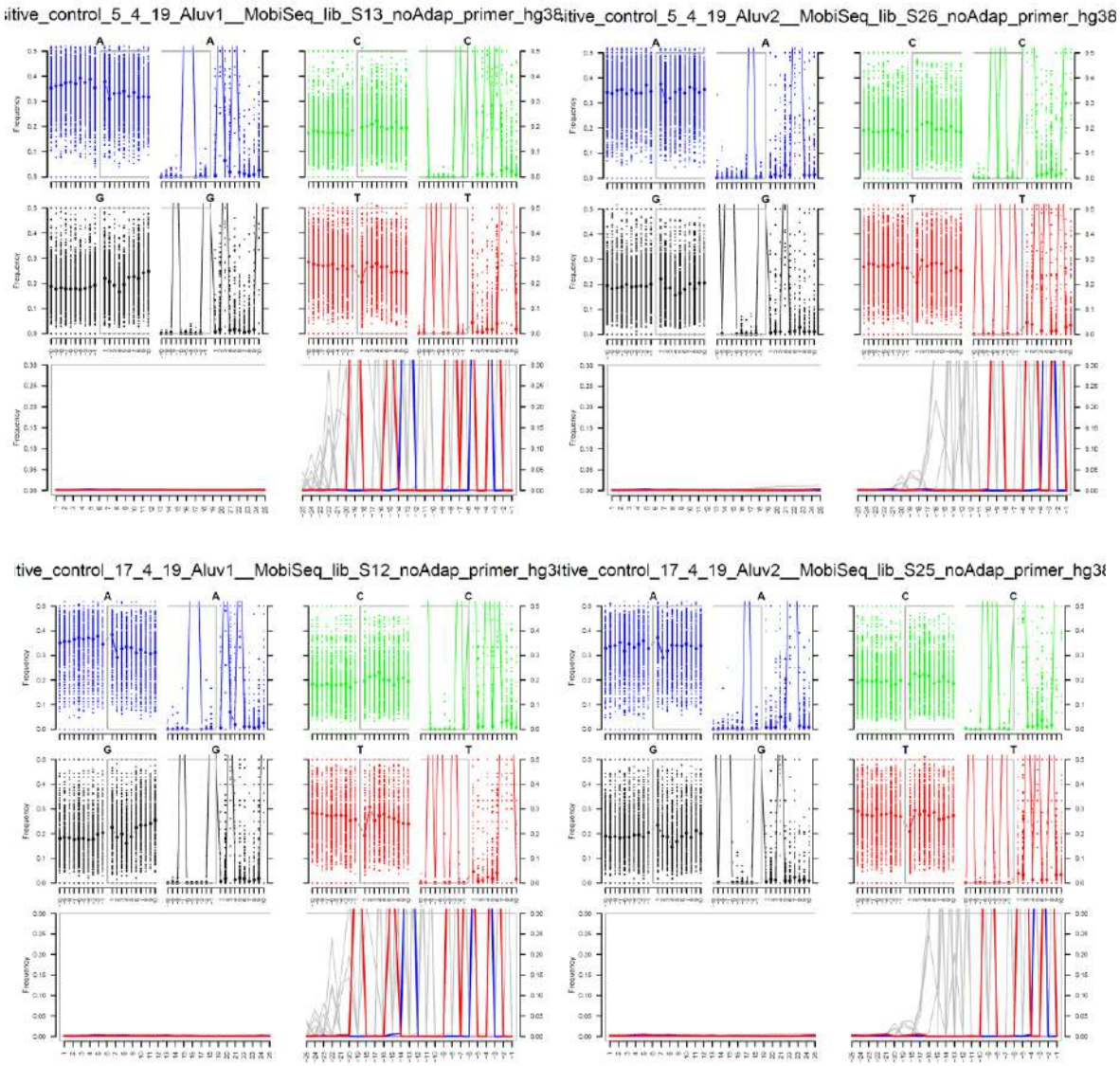
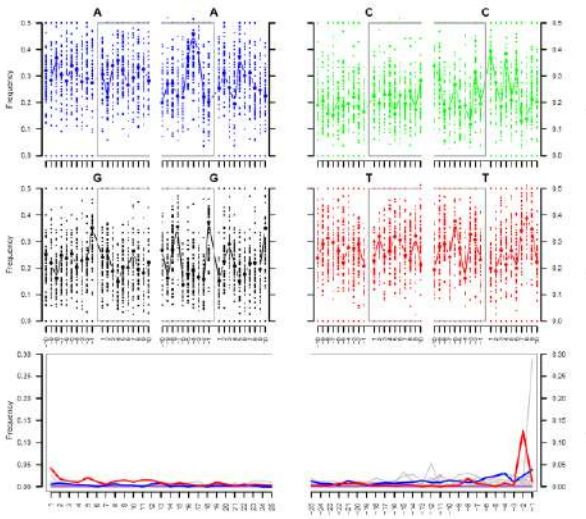
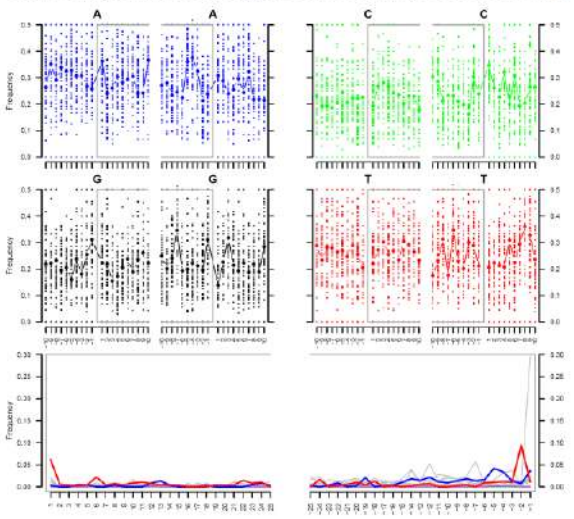


Figure S9. Fragment misincorporation plots for all the MobiSeq libraries.

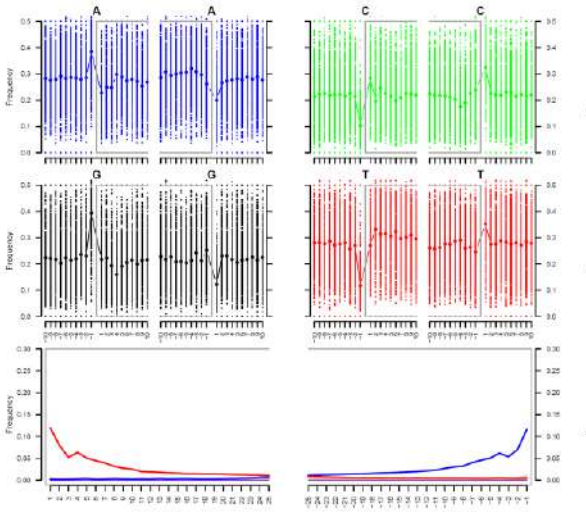
_HVMQ_Extraction_blank_Normal_lib_S47_noAdap_hg38.UCSC.mark



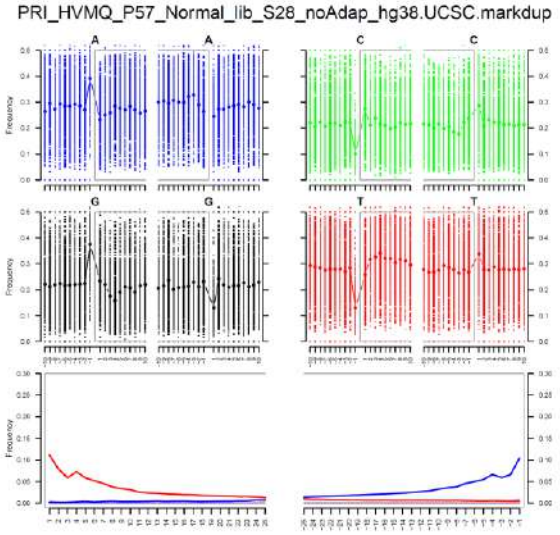
PRI_HVMQ_Blank_Normal_lib_S48_noAdap_hg38.UCSC.markdup



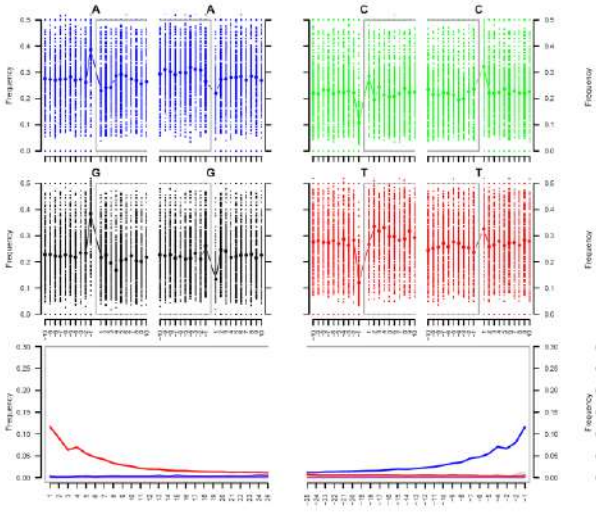
PRI_HVMQ_P56_Normal_lib_S27_noAdap_hg38.UCSC.markdup



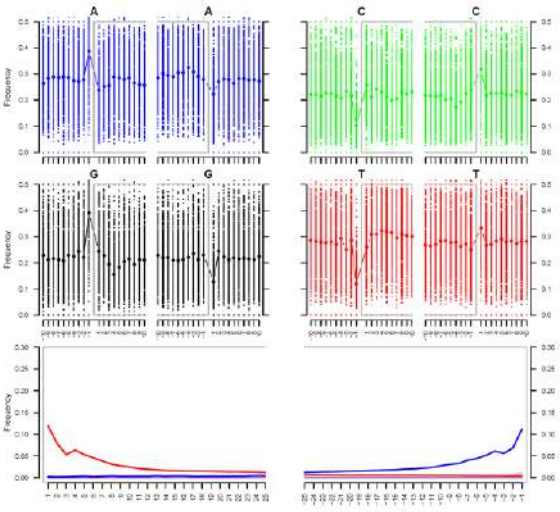
PRI_HVMQ_P57_Normal_lib_S28_noAdap_hg38.UCSC.markdup



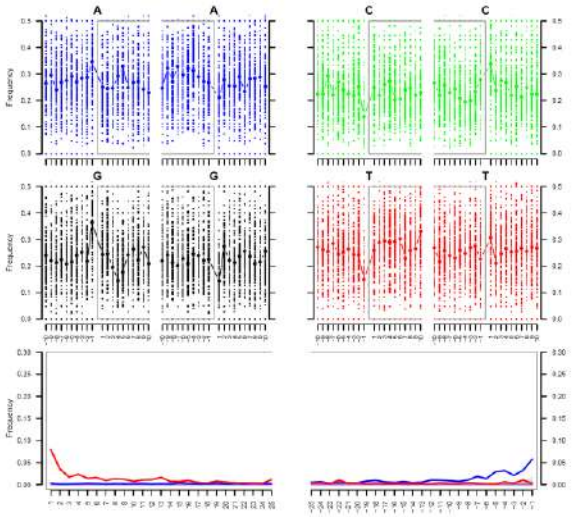
PRI_HVMQ_P58_Normal_lib_S29_noAdap_hg38.UCSC.markdup



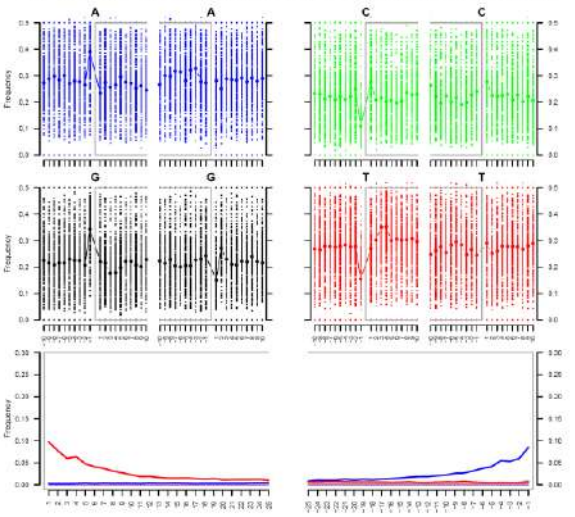
PRI_HVMQ_P59_Normal_lib_S30_noAdap_hg38.UCSC.markdup



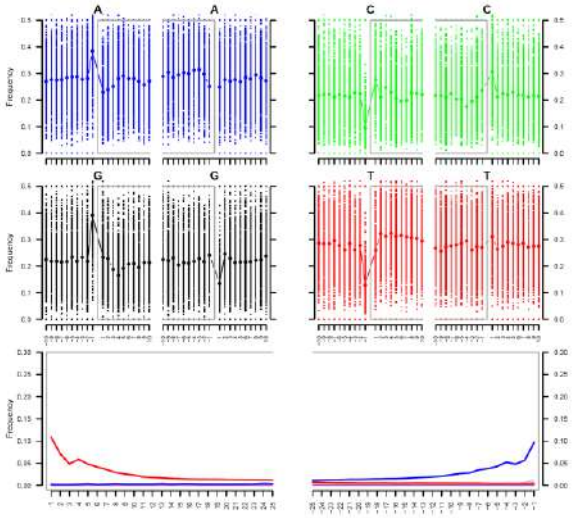
PRI_HVMQ_P60_Normal_lib_S31_noAdap_hg38.UCSC.markdup



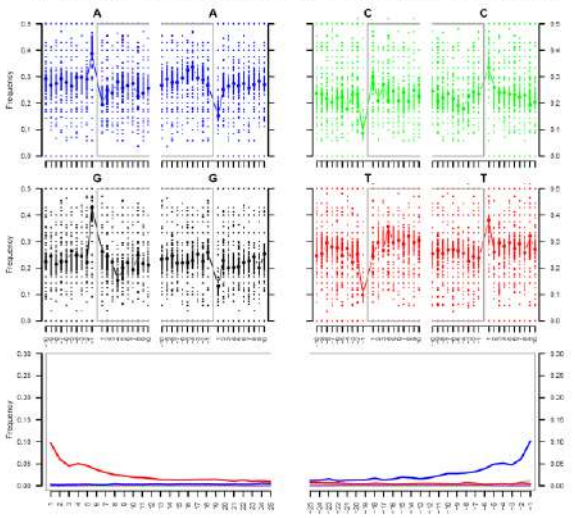
PRI_HVMQ_P61_Normal_lib_S32_noAdap_hg38.UCSC.markdup



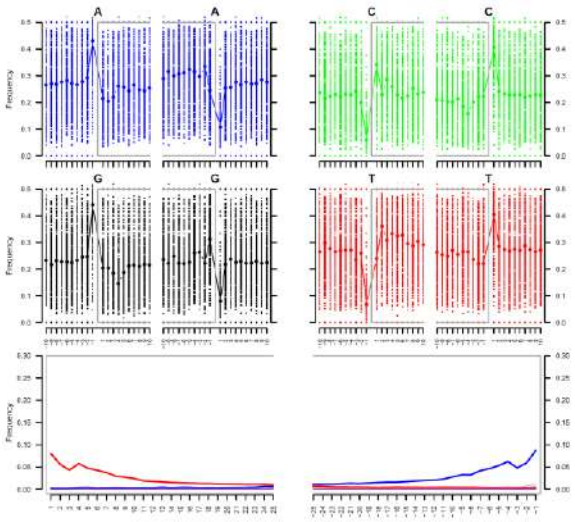
PRI_HVMQ_P62_Normal_lib_S33_noAdap_hg38.UCSC.markdup



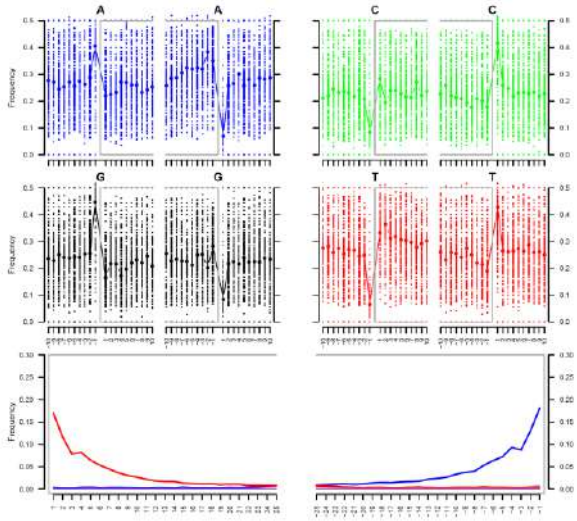
PRI_HVMQ_P63_Normal_lib_S34_noAdap_hg38.UCSC.markdup



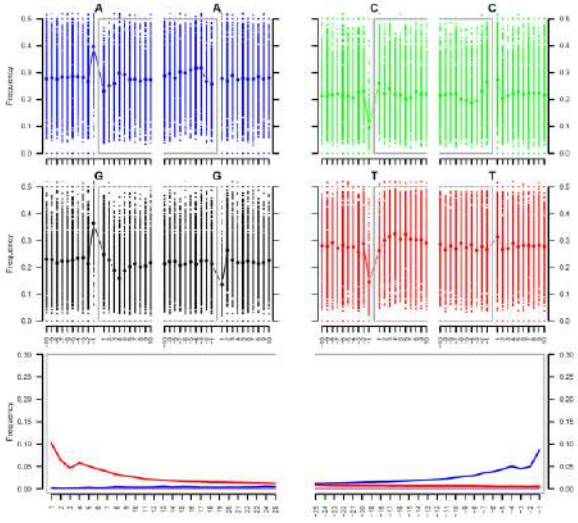
PRI_HVMQ_P64_Normal_lib_S35_noAdap_hg38.UCSC.markdup



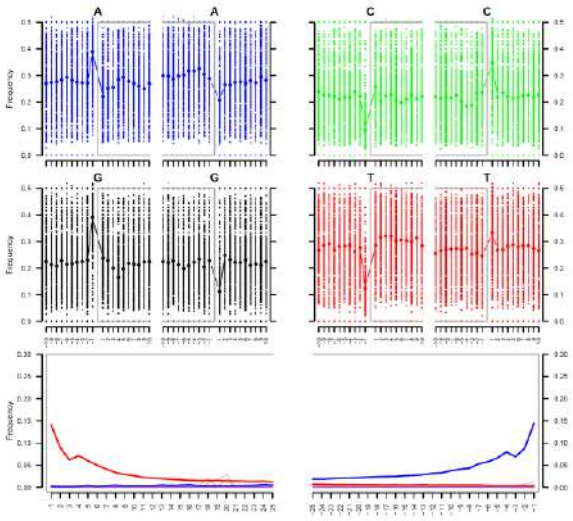
PRI_HVMQ_P65_Normal_lib_S36_noAdap_hg38.UCSC.markdup



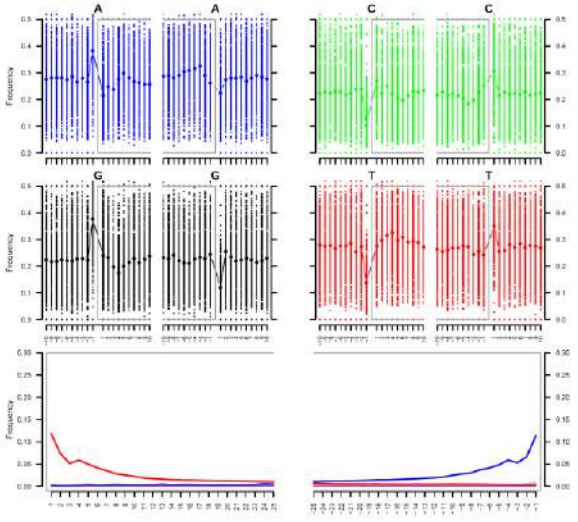
PRI_HVMQ_P66_Normal_lib_S37_noAdap_hg38.UCSC.markdup



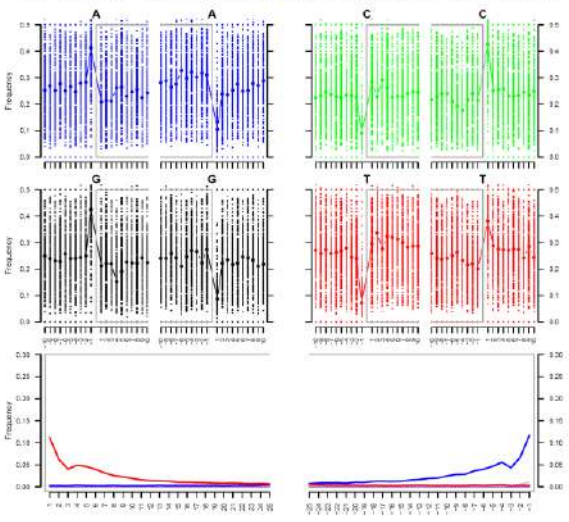
PRI_HVMQ_P67_Normal_lib_S38_noAdap_hg38.UCSC.markdup



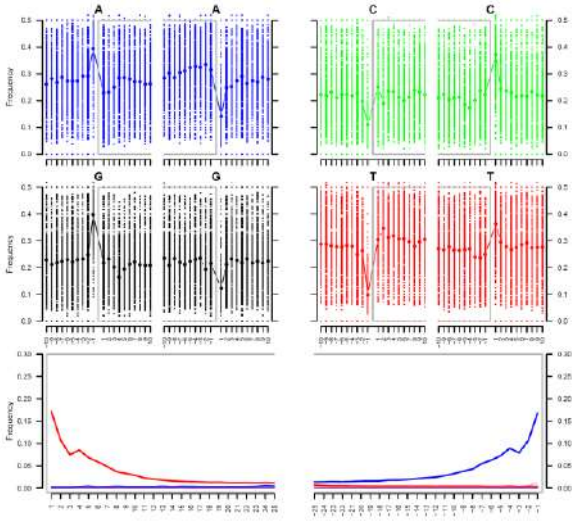
PRI_HVMQ_P68_Normal_lib_S39_noAdap_hg38.UCSC.markdup



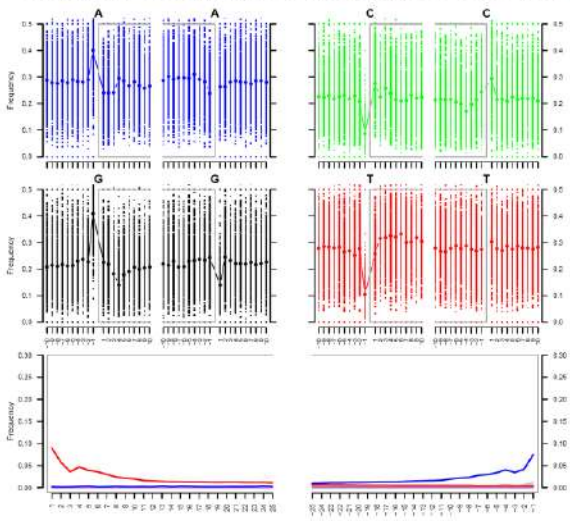
PRI_HVMQ_P69_Normal_lib_S40_noAdap_hg38.UCSC.markdup



PRI_HVMQ_P70_Normal_lib_S41_noAdap_hg38.UCSC.markdup



PRI_HVMQ_P71_Normal_lib_S42_noAdap_hg38.UCSC.markdup



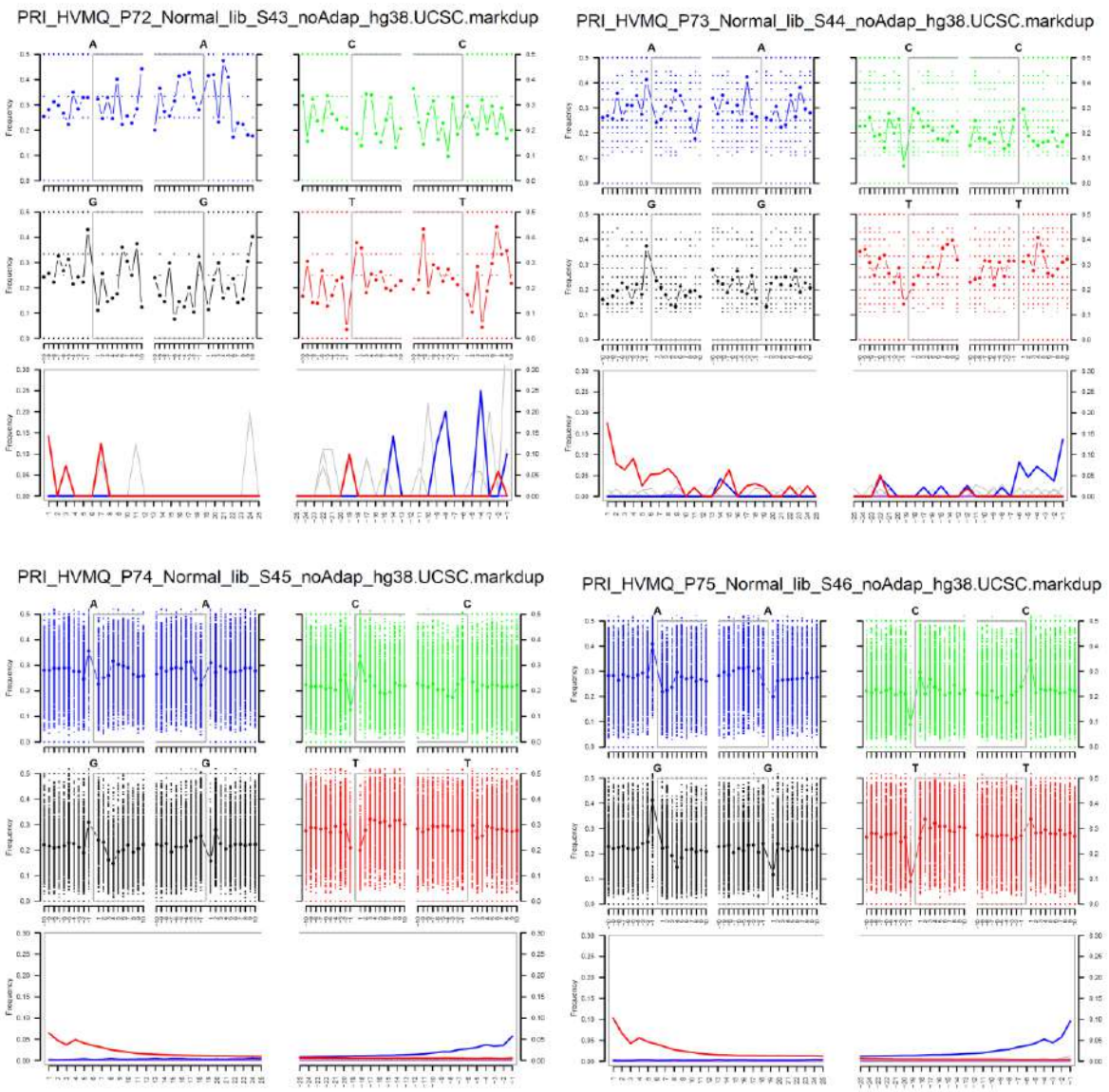
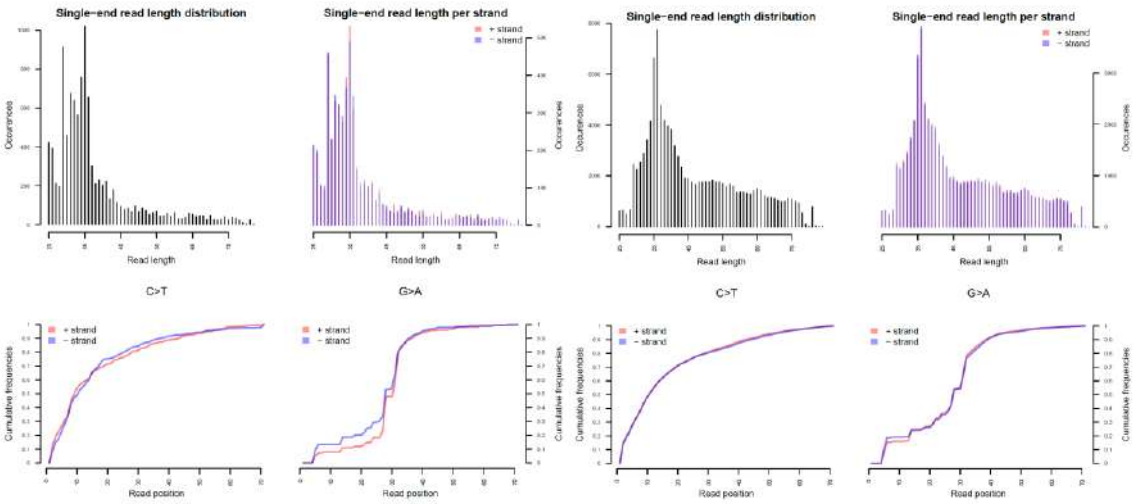


Figure S10. Fragment misincorporation plots for all the shotgun libraries.

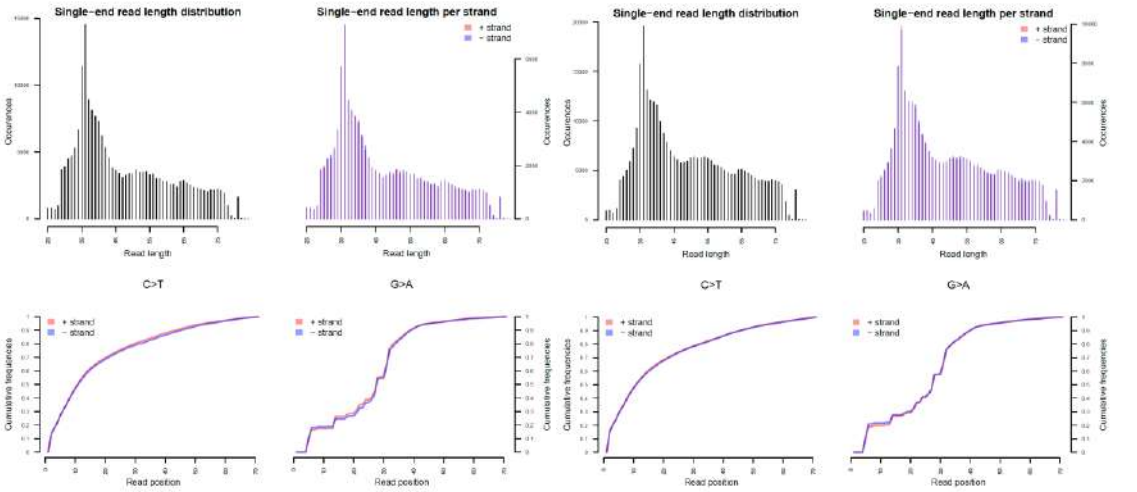
Appendix S5- Length plots

HVMQ_P56_Aluv1__MobiSeq_lib_S1_noAdap_primer_hg38.UCSC.ma



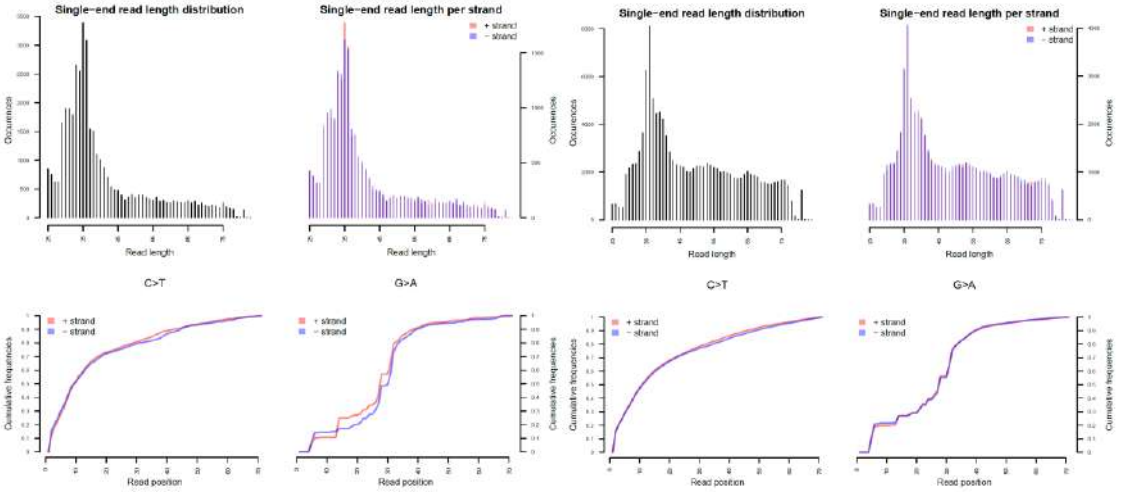
HVMQ_P57_Aluv1__MobiSeq_lib_S2_noAdap_primer_hg38.UCSC.ma

HVMQ_P58_Aluv1__MobiSeq_lib_S3_noAdap_primer_hg38.UCSC.ma



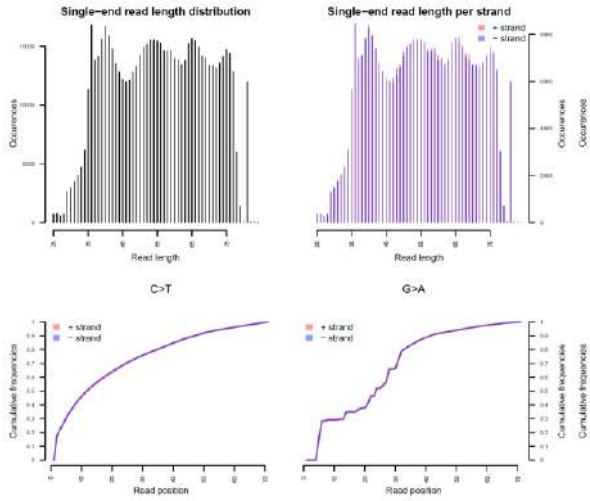
HVMQ_P59_Aluv1__MobiSeq_lib_S4_noAdap_primer_hg38.UCSC.ma

HVMQ_P60_Aluv1__MobiSeq_lib_S5_noAdap_primer_hg38.UCSC.ma

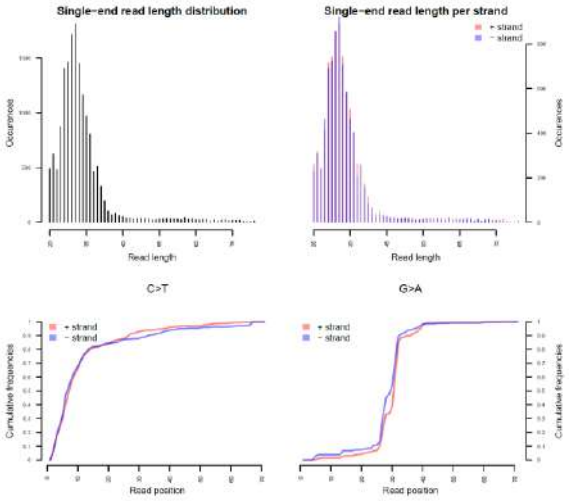


HVMQ_P61_Aluv1__MobiSeq_lib_S6_noAdap_primer_hg38.UCSC.ma

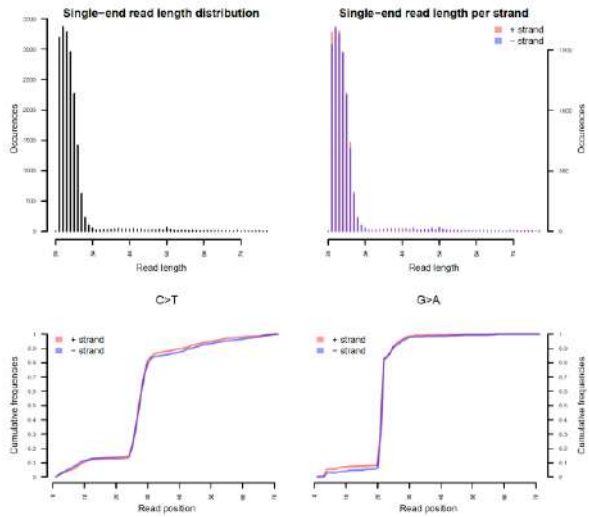
IVMQ_P62_Aluv1__MobiSeq_lib_S7_noAdap_primer_hg38.UCSC.ma



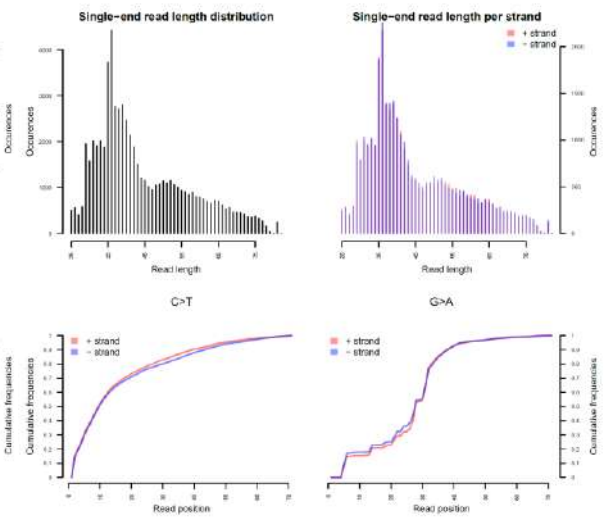
IVMQ_P63_Aluv1__MobiSeq_lib_S8_noAdap_primer_hg38.UCSC.ma



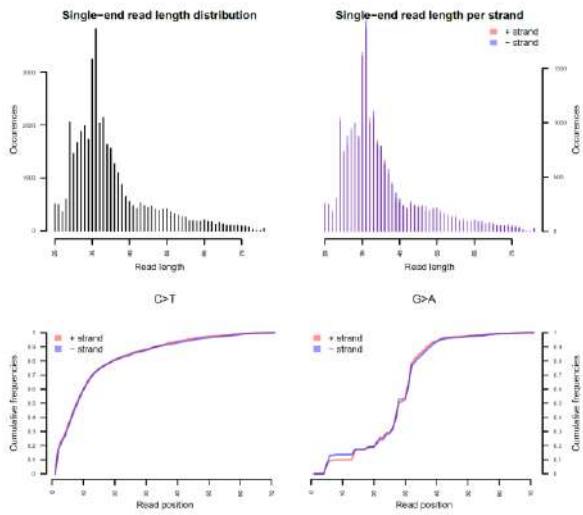
IVMQ_P63_Aluv2__MobiSeq_lib_S24_noAdap_primer_hg38.UCSC.ma



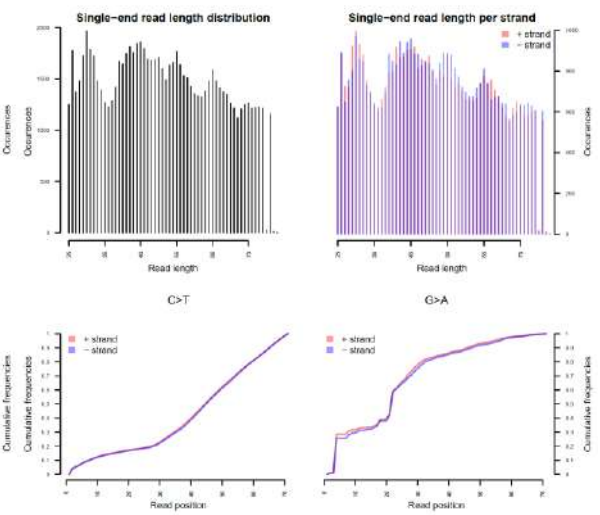
IVMQ_P64_Aluv1__MobiSeq_lib_S9_noAdap_primer_hg38.UCSC.ma



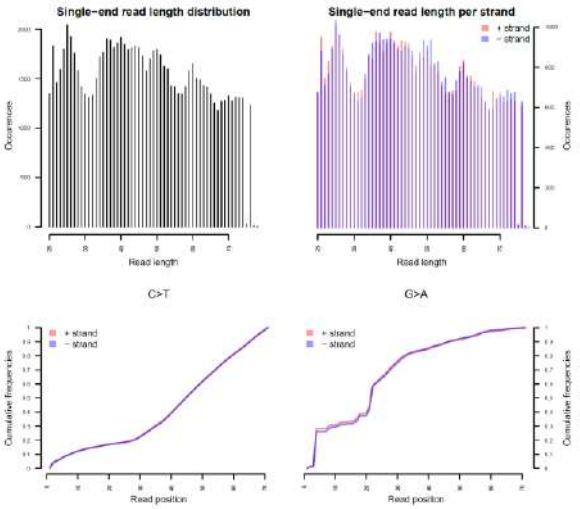
IVMQ_P65_Aluv1__MobiSeq_lib_S10_noAdap_primer_hg38.UCSC.ma



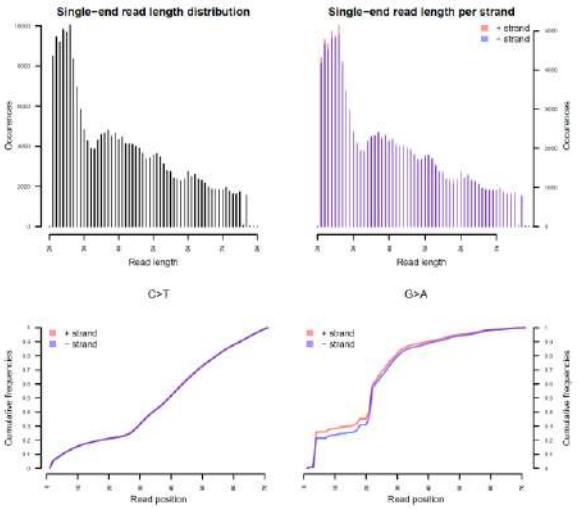
IVMQ_P66_Aluv2__MobiSeq_lib_S14_noAdap_primer_hg38.UCSC.ma



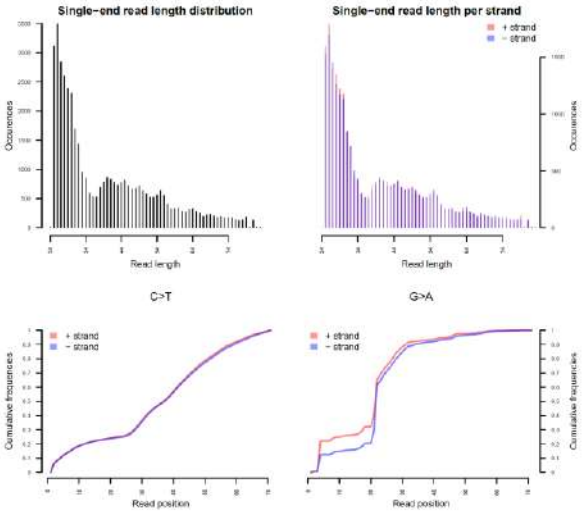
IVMQ_P67_Aluv2__MobiSeq_lib_S15_noAdap_primer_hg38.UCSC.ma



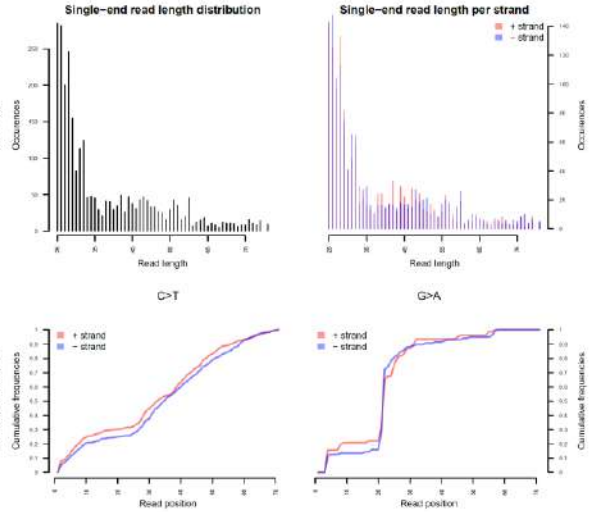
IVMQ_P68_Aluv2__MobiSeq_lib_S16_noAdap_primer_hg38.UCSC.ma



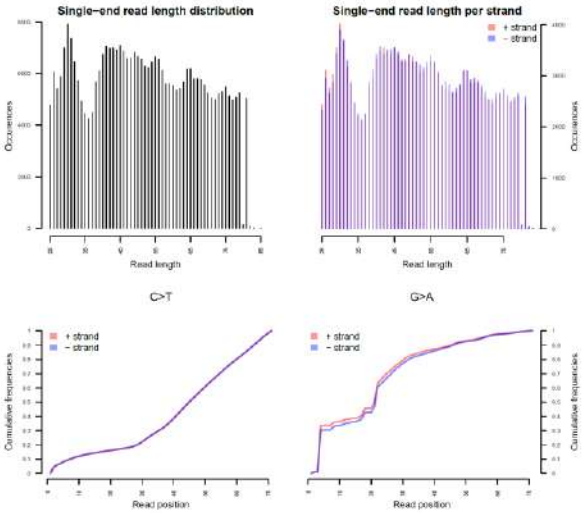
IVMQ_P69_Aluv2__MobiSeq_lib_S17_noAdap_primer_hg38.UCSC.ma



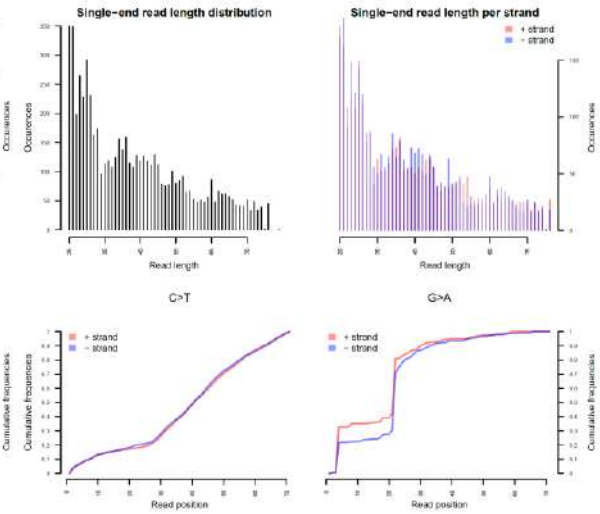
IVMQ_P70_Aluv2__MobiSeq_lib_S18_noAdap_primer_hg38.UCSC.ma



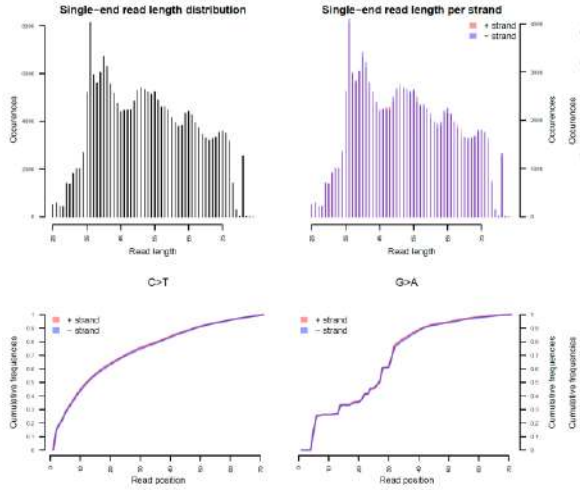
IVMQ_P71_Aluv2__MobiSeq_lib_S19_noAdap_primer_hg38.UCSC.ma



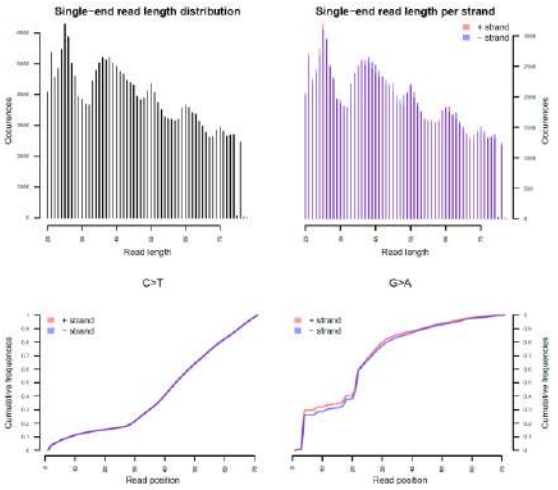
IVMQ_P72_Aluv2__MobiSeq_lib_S20_noAdap_primer_hg38.UCSC.ma



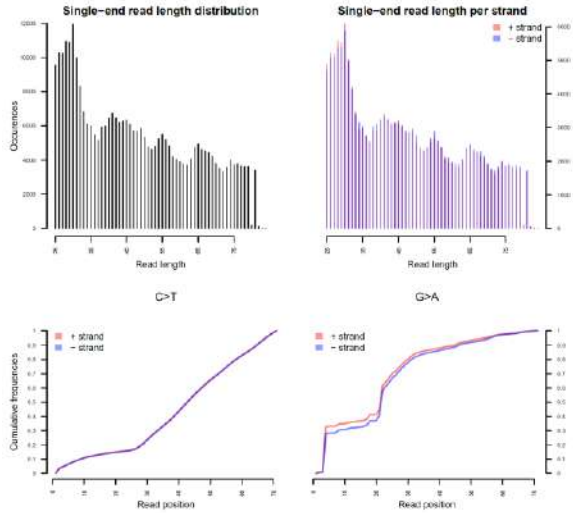
IVMQ_P73_Aluv1__MobiSeq_lib_S11_noAdap_primer_hg38.UCSC.ma



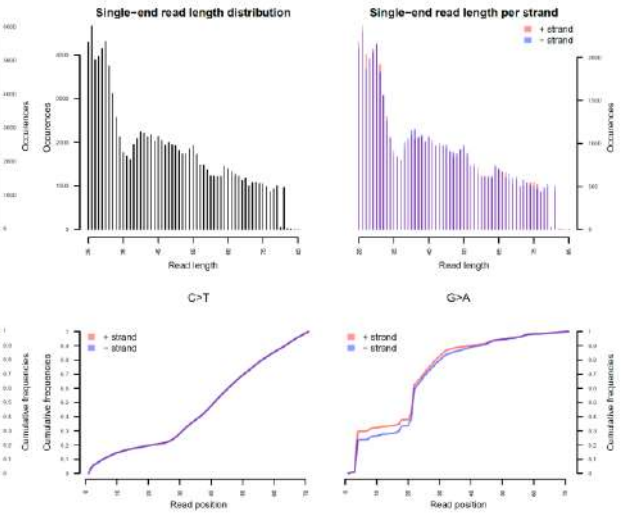
IVMQ_P73_Aluv2__MobiSeq_lib_S21_noAdap_primer_hg38.UCSC.ma



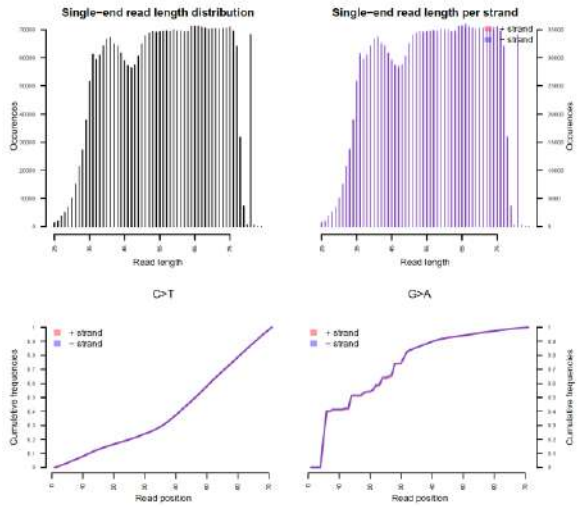
IVMQ_P74_Aluv2__MobiSeq_lib_S22_noAdap_primer_hg38.UCSC.ma



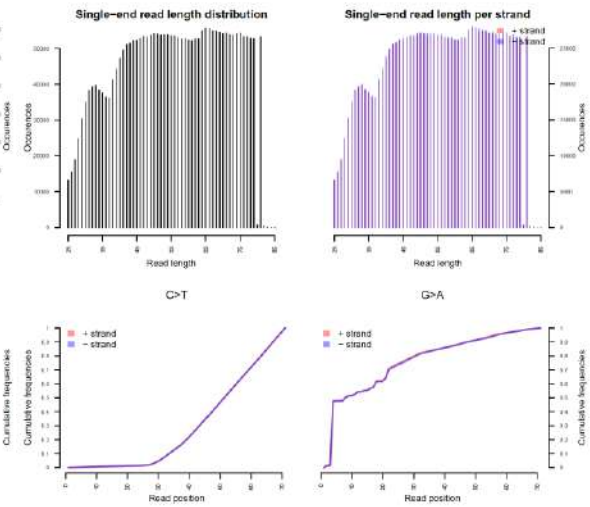
IVMQ_P75_Aluv2__MobiSeq_lib_S23_noAdap_primer_hg38.UCSC.ma



itive_control_5_4_19_Aluv1__MobiSeq_lib_S13_noAdap_primer_hg38



itive_control_5_4_19_Aluv2__MobiSeq_lib_S26_noAdap_primer_hg38



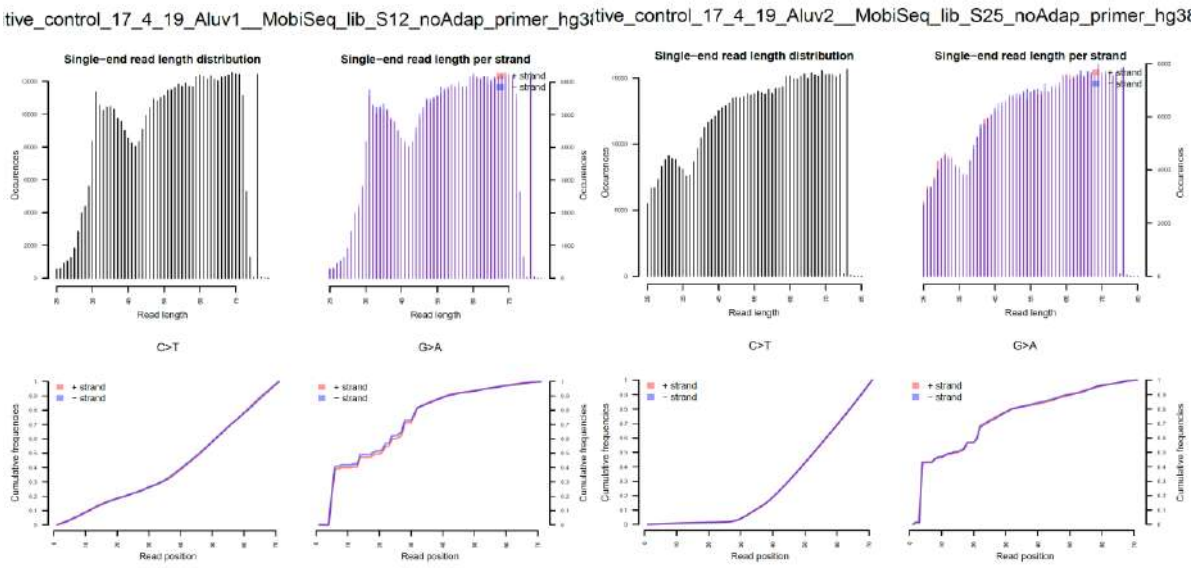
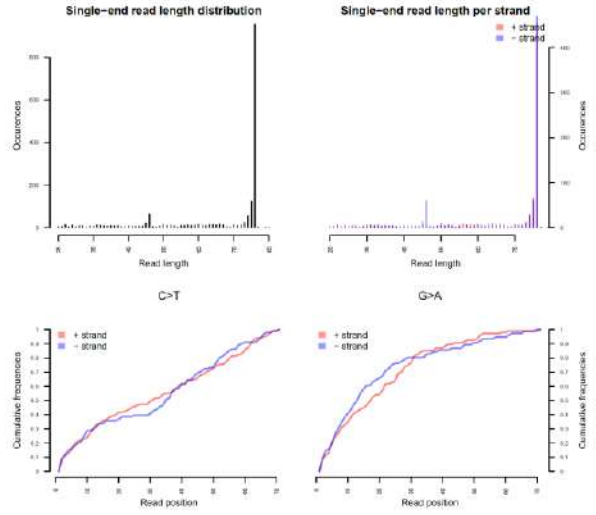
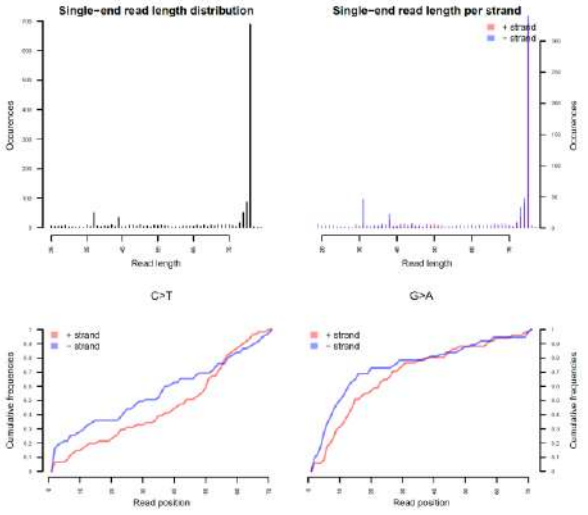


Figure S11. Length plots for all the MobiSeq libraries.

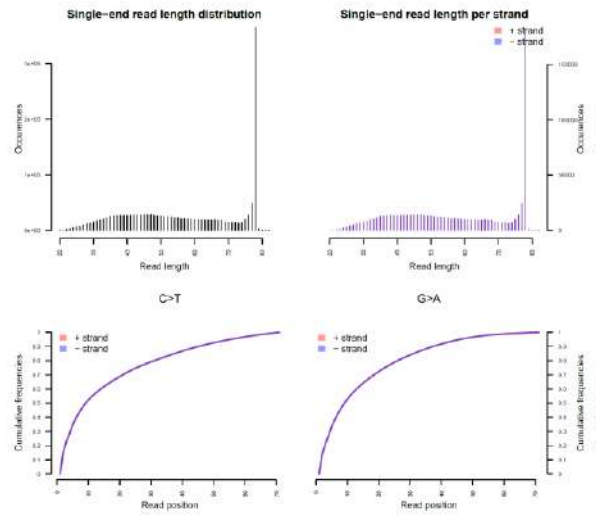
_HVMQ_Extraction_blank_Normal_lib_S47_noAdap_hg38.UCSC.mark



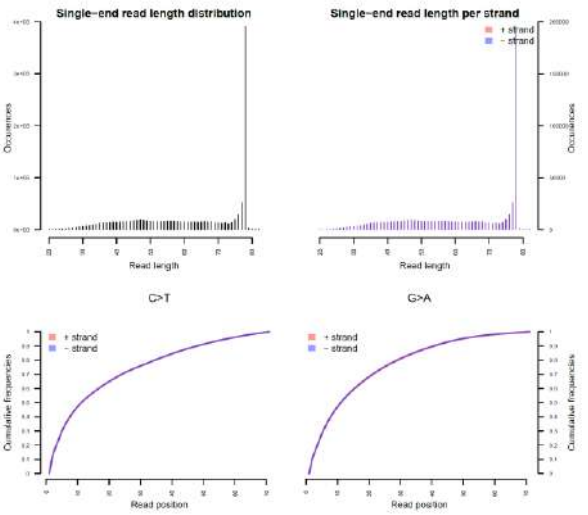
PRI_HVMQ_Blank_Normal_lib_S48_noAdap_hg38.UCSC.markdup



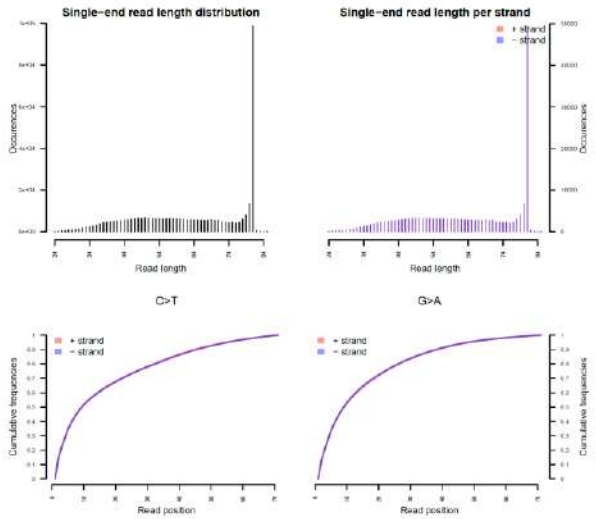
PRI_HVMQ_P56_Normal_lib_S27_noAdap_hg38.UCSC.markdup



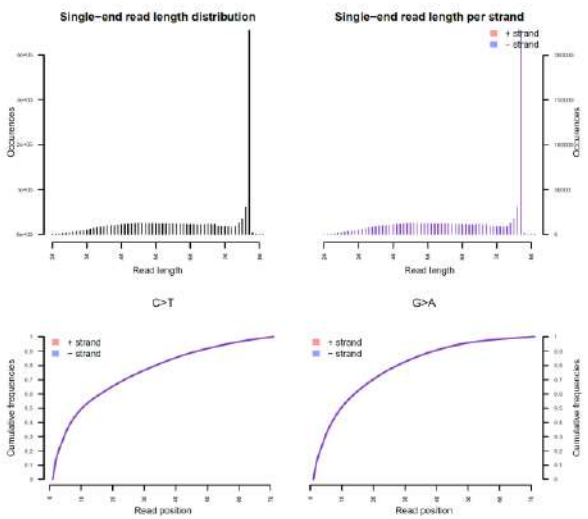
PRI_HVMQ_P57_Normal_lib_S28_noAdap_hg38.UCSC.markdup



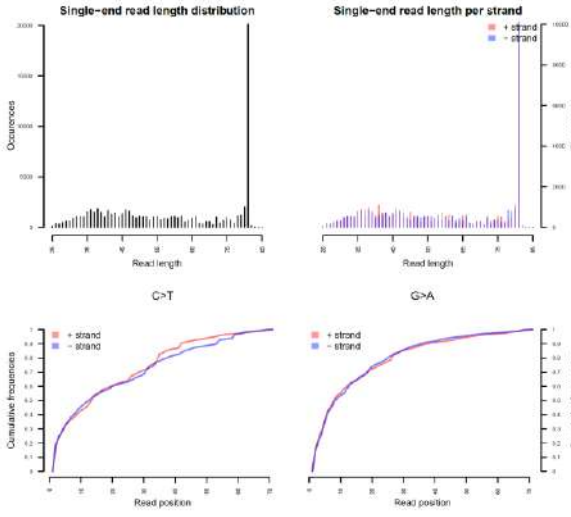
PRI_HVMQ_P58_Normal_lib_S29_noAdap_hg38.UCSC.markdup



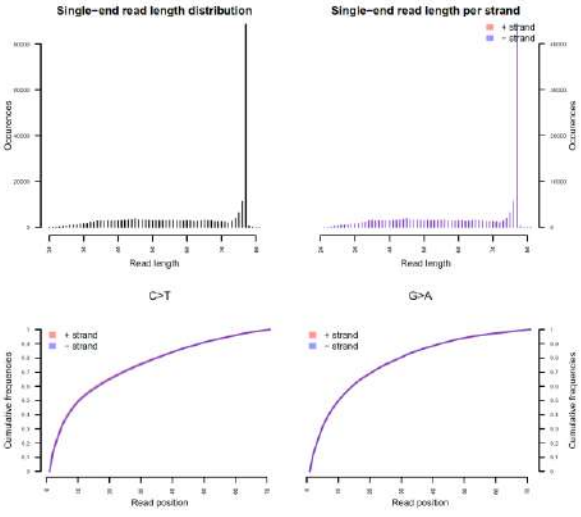
PRI_HVMQ_P59_Normal_lib_S30_noAdap_hg38.UCSC.markdup



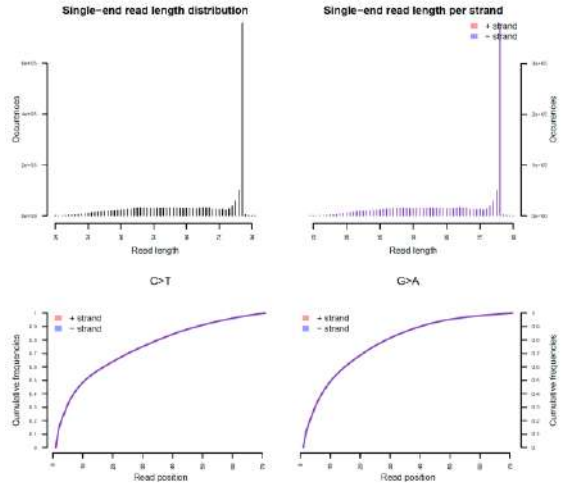
PRI_HVMQ_P60_Normal_lib_S31_noAdap_hg38.UCSC.markdup



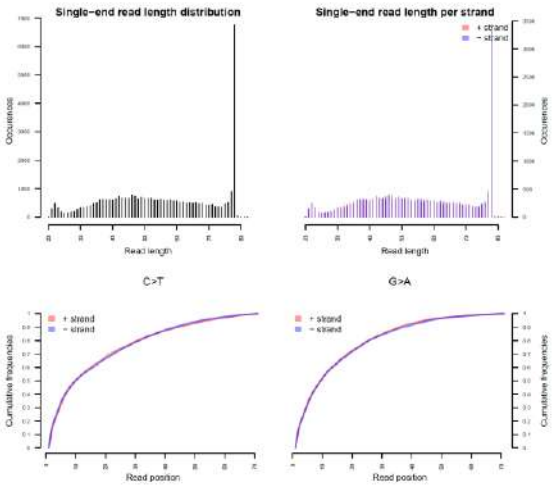
PRI_HVMQ_P61_Normal_lib_S32_noAdap_hg38.UCSC.markdup



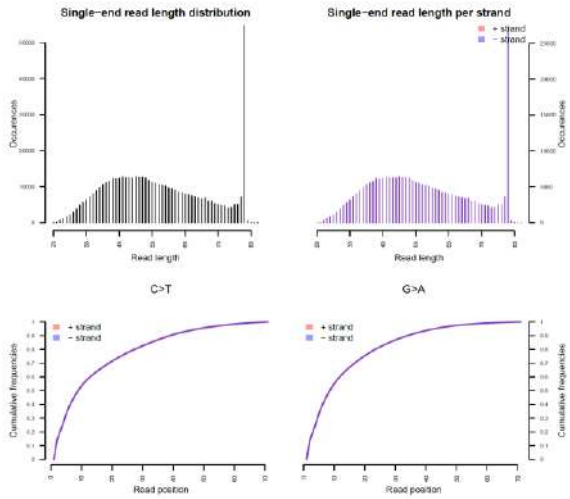
PRI_HVMQ_P62_Normal_lib_S33_noAdap_hg38.UCSC.markdup



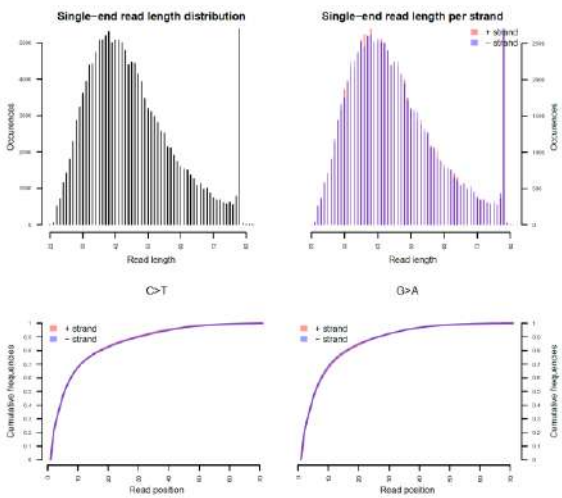
PRI_HVMQ_P63_Normal_lib_S34_noAdap_hg38.UCSC.markdup



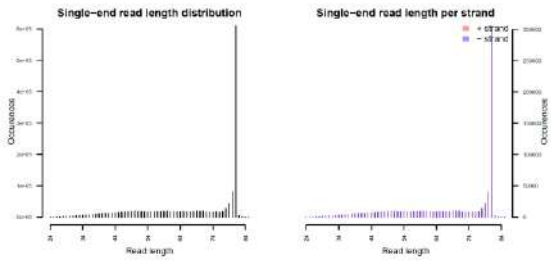
PRI_HVMQ_P64_Normal_lib_S35_noAdap_hg38.UCSC.markdup



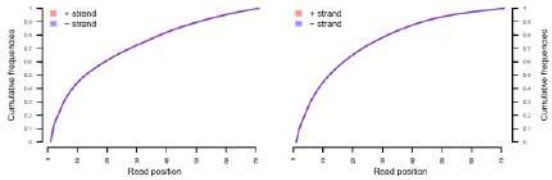
PRI_HVMQ_P65_Normal_lib_S36_noAdap_hg38.UCSC.markdup



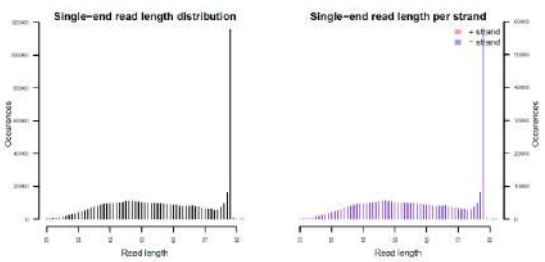
PRI_HVMQ_P66_Normal_lib_S37_noAdap_hg38.UCSC.markdup



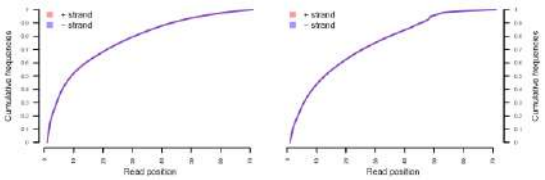
C>T G>A



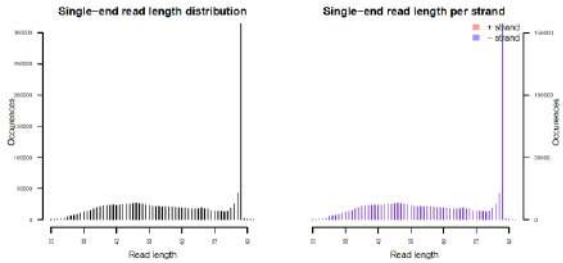
PRI_HVMQ_P67_Normal_lib_S38_noAdap_hg38.UCSC.markdup



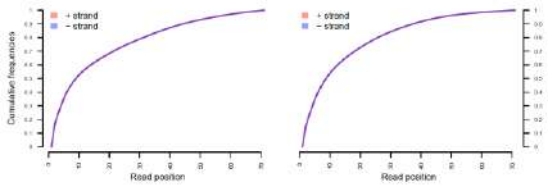
C>T G>A



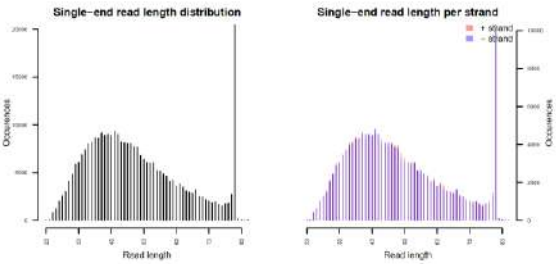
PRI_HVMQ_P68_Normal_lib_S39_noAdap_hg38.UCSC.markdup



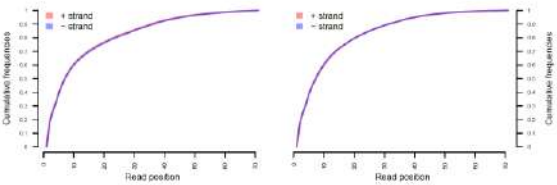
C>T G>A



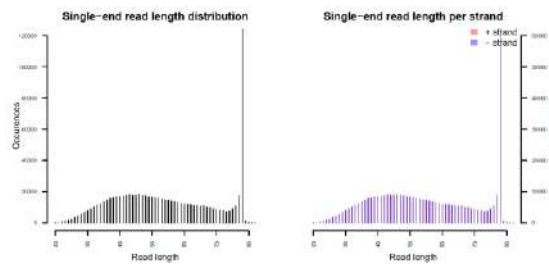
PRI_HVMQ_P69_Normal_lib_S40_noAdap_hg38.UCSC.markdup



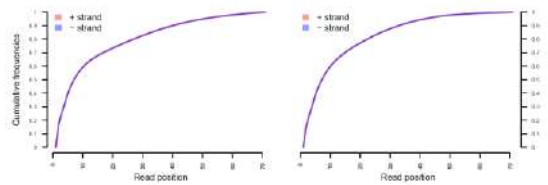
C>T G>A



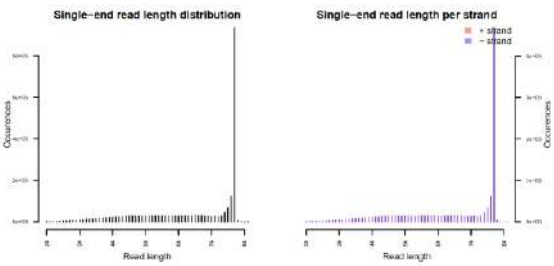
PRI_HVMQ_P70_Normal_lib_S41_noAdap_hg38.UCSC.markdup



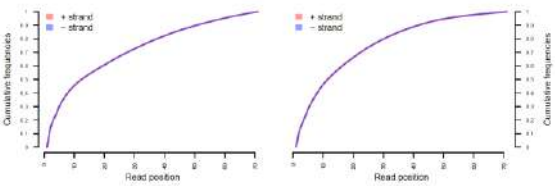
C>T G>A



PRI_HVMQ_P71_Normal_lib_S42_noAdap_hg38.UCSC.markdup



C>T G>A



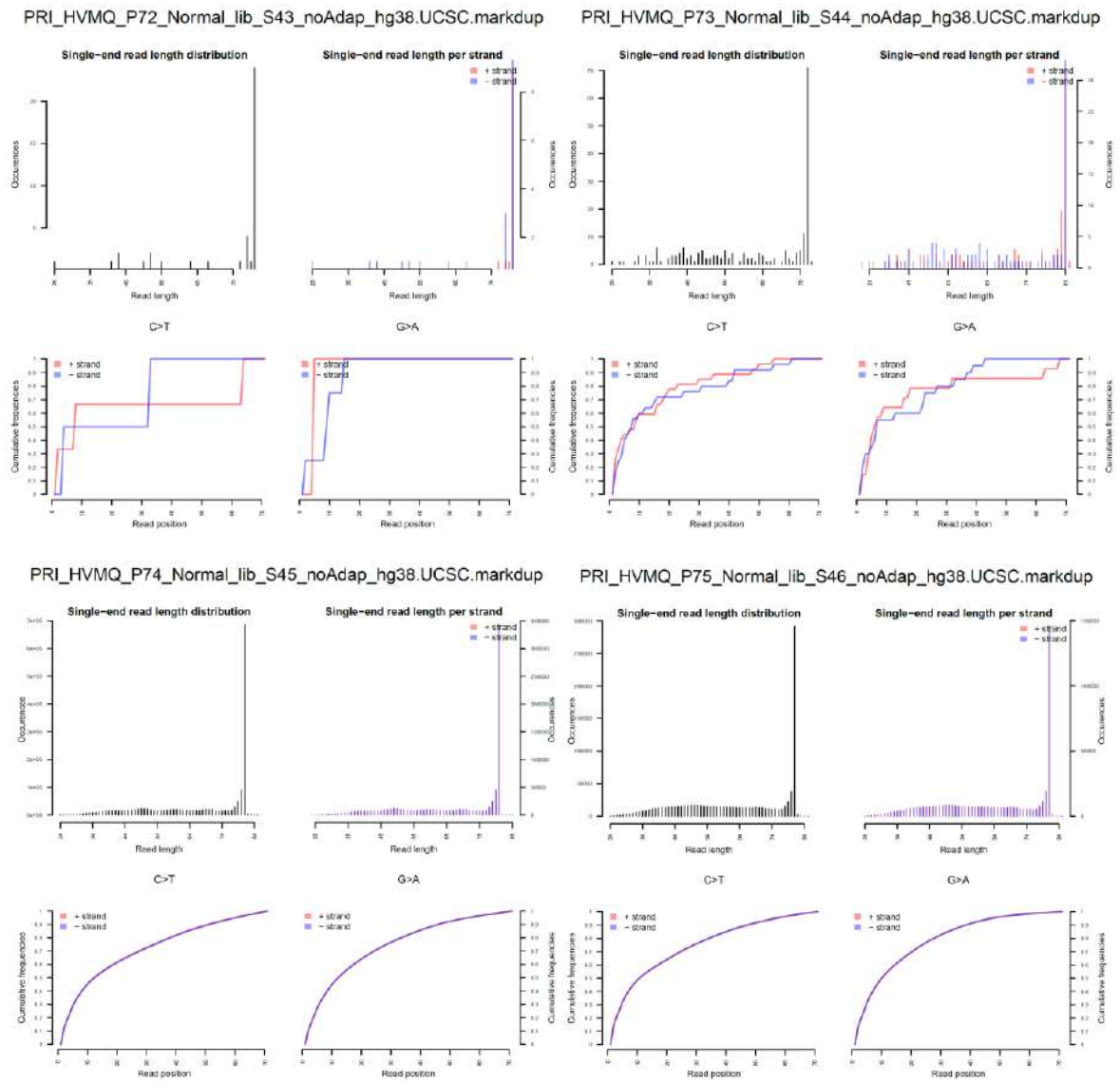


Figure S12. Length plots for all the shotgun libraries.