UNIVERSITY OF CALIFORNIA
SANTA CRUZ

METHODS DEVELOPMENT FOR GENOME SEQUENCING APPLICATIONS

A dissertation submitted in partial satisfaction
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Balaji Sundararaman

June 2023

The Dissertation of Balaji Sundararaman
is approved:

____
Professor Richard Ed Green, Chair

____
Professor Russell B Corbett-Detig

____
Professor Mark Akeson

____
Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

METHODS DEVELOPMENT FOR GENOME SEQUENCING APPLICATIONS

by

Balaji Sundararaman

Degraded DNA isolated from diverse samples like ancient bones, environmental samples, forensic specimens and from blood plasma hold a wealth of information. However, due to the quality and quantity of the DNA isolated from these samples render them difficult to analyze using shotgun sequencing. Enrichment of the target regions of interest instead of sequencing the entire DNA is a cost-effective method. However, DNA/RNA baits need to enrich the target regions are expensive. In this thesis, I present a method for cost-effective DNA bait synthesis that I named as *C*ircular *N*ucleic acid *E*nrichment *R*eagent (CNER, pronounced as *snare*) synthesis method. I demonstrate the application of the CNER method to make probes for specific target regions and for whole-genome enrichment (WGE). First, I use the CNER method to make probes for targeted genotyping of ~23k SNPs in the horse genome, using which I studied the demographic history of Late Pleistocene horses. Next, I demonstrate the CNER method to make WGE probes to detect and enrich entire genomes of Tuberculosis causing bacteria and Toxoplasmosis causing parasite. Finally, I demonstrate the CNER method to generate probes to enrich ~108k SNP markers for genotyping DNA isolated from rootless hair for forensic application.

I dedicate this work to my Appa, Amma, Divya and Papa.

Acknowledgements

I sincerely thank my advisors Prof. Ed Green and Prof. Beth Shapiro for providing me the great opportunity to work in the UCSC Paleogenomics lab. Beth and Ed, I am indebted to you for your trust in my abilities, for providing me the space and time to learn new avenues, for letting me to explore on my own, and for the support and guidance when I needed them the most. Thank you also for introducing me to the ancient DNA world. I am grateful to you for supporting and guiding my entrepreneurial endeavors.

I am thankful to my thesis committee member, Prof. Mark Akeson for his constructive criticism of my research. Mark, thank you for writing letter of recommendation to secure the NIJ GRF fellowship and for sharing your great insights on entrepreneurship.

I am grateful to my thesis committee member, Prof. Russell Corbett-Detig for being my grad school life coach. Russ, thank you for pushing me to think big about the applications of CNERs technology and for introducing me to the CDPH collaboration.

I am grateful to Dr. Matthew Sylvester, Dr. Varvara Kozyreva and Dr. Zenda L. Berrada of the California Department of Public Health for re-kindling my interest in Tuberculosis. I really appreciate their generous gift of TB DNA, guidance, review and feedback on the TB project.

I am thankful to Dr. Karen Shapiro of UC Davis for introducing and teaching me about Toxoplasmosis. I am also thankful to the members of her lab at UC Davis for providing the parasite DNA for my research.

I am grateful to the members of UCSC Paleogenomics lab circa 2018-2023 with whom I had the great pleasure working with. Specially, I thank Dr. Rachel Meyer for her great enthusiasm on my projects and for introducing me to new collaborators. I had great time working with Alisa Vershinina and Samantha Hershauer on the ancient horse project. I thank Josh Kapp for all the great conversations over pipetting on molecular bio tech-devs. I am thankful to the PGL Sequencing team especially Nick Maurer, Molly Cassatt, Shelby Dunn, Sam Cutler, Sam Sacco, Ciara Wanket, Will Seligman and Sarah Ford for helping on my sequencing needs. I am also thankful to Merly Escalona, Nathan Schaefer, Jonas Oppenheimer and Remy Nguyen for being my tech support and for being friendly ‘shark tank’ residents.

I thank my classmates and colleagues in the 2017 and 2016 UCSC BME cohort for being great friends who made my grad school life such a sense of community. Thank you for the great times and memories.

I am thankful to Mr. Jeff Jackson, Director of Intellectual Property Management for going extra mile to educate me the details of IP processes and for kindling my interest beyond academia. I also thank Mr. Ryan Sharp, Mr. Michael Alvarez, Ms. Ann Pace of the Office of Research and UCSC QB3 program and the inaugural UCSC - Nucleate Innovation Fellowship for nurturing my entrepreneurial interest.

I am grateful to the Department of Justice, National Institute of Justice Graduate Research Fellowship (DOJ-NIJ GRF) grant number 2020-R2-CX-0029 for supporting my research. This grant supported both my forensic and non-forensic projects. I thank Ms. Cathy Girouard, DOJ Program Manager and members of the UCSC Office of Sponsored Projects for administrating this grant.

I am grateful to the UCSC Office of Research, Industry Alliances & Technology Commercialization (IATC) 2019 Launchpad Proof-of-Concept Grant and the California State Bill AB2664 for providing the much-needed initial financial support for proof-of-concept experiments.

I am grateful to Mr. Mothusi Pahl and Mr. Andrea Carafa, UCSC-CIED, the judges and sponsors of the UCSC Slug Tank 2019 and IdeaHub 2021 for supporting my business idea out of this thesis research.

I thank the UCSC Slug Support and STARS programs, and the FSH Community for their financial and emotional support especially during the tough times of 2020.

I am indebted to my sisters, nephews, nieces, uncles and in-laws for their love, care and support in all my endeavors.

Introduction

Degraded DNA isolated from diverse samples like ancient samples, environmental samples, forensic specimens and from blood plasma hold a wealth of information. However, degraded DNA isolated from ancient samples know as ancient DNA (aDNA), from environmental samples known as environmental DNA (eDNA), from forensic specimens known as forensic trace DNA (ftDNA) and from cell free/circulating tumor DNA (cf/ctDNA) exhibit three characteristics that impede their analyses. First, DNA recovery from these samples is challenging and yields low quantities due to sample availability and DNA content. Second, the recovered DNA is of low quality, degraded and chemically modified making it incompatible or inefficient for typical methods. Finally, these samples have high levels of unwanted DNA contamination. Research in the past decade facilitated massively parallel sequencing (MPS) also known as next generation sequencing (NGS) methods to recover sequence information from degraded DNA. Protocols have been developed to convert sparse and degraded DNA into sequenceable libraries by ligating universal adapters. However, inadvertent sequencing of unwanted DNA exhausts sequencing resources.

Enrichment of the target DNA of interest for sequencing is a cost-effective method that also improves sensitivity and specificity than shotgun sequencing. Targeted sequencing of regions of interest instead of the whole genome is used to identify rare variants. Targeted sequencing of the genes frequently mutated in cancer is widely used for companion diagnostics using ct/cfDNA. Targeted sequencing of select single nucleotide polymorphic (SNP) sites in the genome is used for analyses of aDNA to study population demography of extinct and extant species. Targeted genotyping by sequencing (GBS) of SNPs is also used for ftDNA analyses to solve cold cases and for missing person and victim identification. PCR-based target enrichment methods often fail for degraded DNA due to amplification failure caused by PCR inhibition or degradation of priming sites.

In-solution hybridization capture methods for target enrichment use biotinylated DNA or RNA molecules called 'baits'. These baits are complementary to the target DNA sequences, hence form heteroduplex which are enriched on streptavidin magnetic beads and the non-specific DNA is washed away. RNA baits are generated by in-vitro transcription of synthetic DNA oligonucleotides and genomic DNA fragments. These methods use one or more steps of oligo synthesis, DNA ligation, PCR amplification, and in vitro transcription of target sequences, all of which are known to cause biases. DNA or RNA baits synthesis for targeted sequencing requires solid-phase oligonucleotide synthesis and/or in vitro transcription, both methods have drawbacks like incomplete chemical synthesis of the ends of long oligos and expensive large-scale synthesis.

To overcome the disadvantages of current bait synthesis methods for in-solution hybridization capture target enrichment methods, I developed a cost-effective DNA bait synthesis method for specific target regions or for whole-genome enrichment (WGE) that I named as Circular Nucleic acid Enrichment Reagent synthesis (CNERs, pronounced as *snares*). In the first chapter, I used the CNER method to make probes for targeted genotyping of ~23k SNP sites in the horse genome to study the demographic history of horses in the Beringia region. For the same set of SNP markers, we also purchased RNA baits from a commercial vendor. I compared the performance of the CNERs with the commercial RNA baits for analyzing aDNA isolated from ten Late Pleistocene horse bone samples collected in the Beringia region. I showed that CNERs are about two-fold more efficient in SNP enrichments that results in larger number of targeted SNPs recovered with higher coverage compared to the commercial RNA baits. I also demonstrated that the data generated by CNERs and RNA baits results in identical genotypes and population structure.

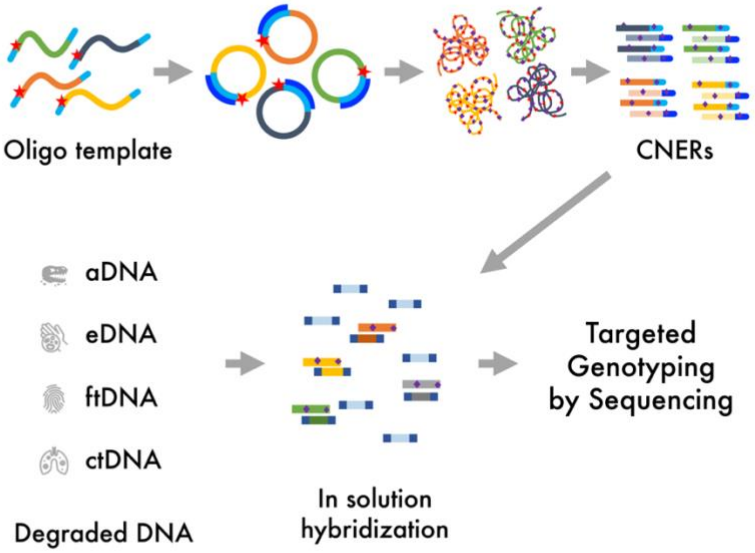


Figure 0.1 Overview of the CNERs method for targeted genotyping of degraded DNA isolated from diverse samples.

In the second chapter, I used the CNER method to make WGE probes to enrich entire genomes of Tuberculosis (TB) causing bacteria, *Mycobacterium tuberculosis (M. tuberculosis)*. *M. tuberculosis* is a slow growing bacterium that also develops resistance to antibiotic used to treat TB. I showed that the CNERs-WGE method can detect up 100 genome copies of *M. tuberculosis* DNA spiked in against vast amount of human DNA background. Using an existing pipeline, I demonstrate that the CNERs-WGE data can be used to identify lineages and drug-resistance patterns. Further, in the second chapter, I also showed the CNERs-WGE method to enrich and detect *Toxoplasma gondii*, a universal parasite that can cause severe disease in immunocompromised individuals and spread through contaminated food and water.

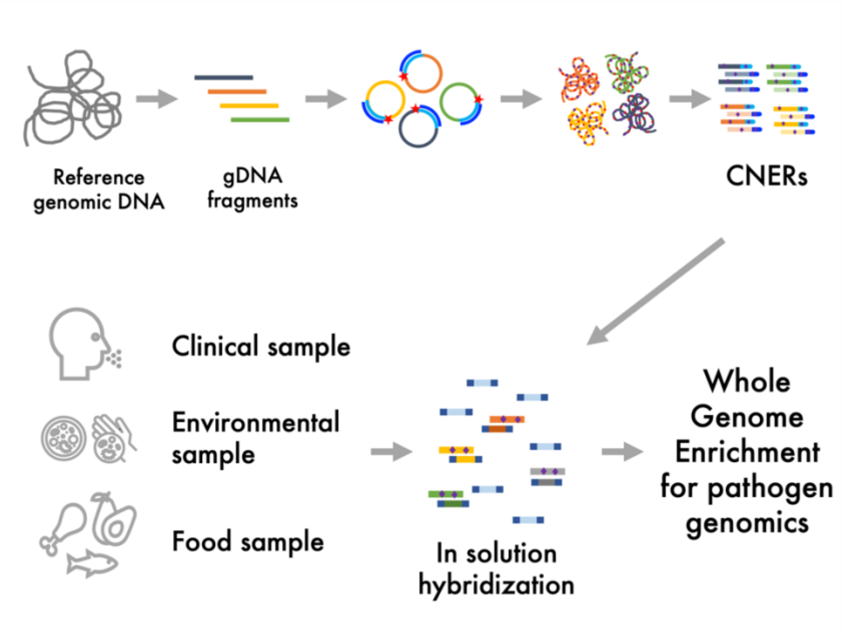


Figure 0.2 Overview of the CNERs Whole Genome Enrichment method for pathogen genomics.

In third chapter, I demonstrated the CNERs method to generate ~108k SNPs markers common to three major direct-to-consumer (DTC) genetic testing platforms. I optimized a subset of 36k SNP makers for genotyping DNA isolated from rootless hair samples collected from 50 volunteers. I compared the genotypes from the CNERs data with the whole genome sequencing (WGS) data generated using DNA isolated from saliva samples from the same individuals. I discuss the application of these CNERs panels for genetic genealogy searches to solve cold cases.

Chapter 1. A method to generate capture baits for targeted sequencing

Abstract

Hybridization capture approaches allow targeted high-throughput sequencing analysis at reduced costs compared to shotgun sequencing. Hybridization capture is particularly useful in analyses of genomic data from ancient, environmental, and forensic samples, where target content is low, DNA is fragmented and multiplex PCR or other targeted approaches often fail. Here, we describe a DNA bait synthesis approach for hybridization capture that we call *C*ircular *N*ucleic acid *E*nrichment *R*eagent, or CNER (pronounced "*snare"*). The CNER method uses rolling-circle amplification followed by restriction digestion to discretize microgram quantities of hybridization probes. We demonstrate the utility of the CNER method by generating probes for a panel of 23,771 known sites of single nucleotide polymorphism in the horse genome. Using these probes, we capture and sequence from a panel of ten ancient horse DNA libraries, comparing CNER capture efficiency to a commercially available approach. With about one million read pairs per sample, CNERs captured more targets (90.5% versus 66.5%) at greater mean depth than an alternative commercial approach.

Introduction

Compared with whole-genome sequencing, targeted sequencing is a cost-effective method for analyzing specific genomic regions (1). Targeted sequencing has wide application in diagnostics, metagenomic, phylogenetic, ancient and environmental DNA studies, and forensics (2, 3). In targeted sequencing, regions of interest are enriched by hybridization capture using target-specific probes or by PCR amplification using target-specific primers, followed by high-throughput next-generation sequencing (NGS). Hybridization capture methods overcome drawbacks of PCR-based target enrichment, including scalability to a large number of targets, PCR failure, and PCR artifacts (1, 2).

Pioneering hybridization capture experiments used DNA arrays to enrich for targeted sequencing of human samples (4–7) and Neanderthal ancient DNA (aDNA) (8). In these array-based hybridization capture methods, NGS library molecules were hybridized to a microarray imprinted with probes targeting human exons. After washing non-hybridized library molecules off the surface of the array, captured molecules were eluted and sequenced (4–8). Array-based hybridization capture expanded the capability to millions of target regions, beyond what is achievable with PCR-based enrichment methods (1–3). However, array-based capture is labor and time-intensive and requires large amounts of input DNA as well as specialized instrumentation for capture.

In-solution hybridization capture is currently the most commonly used method of targeted sequencing due to the commercial availability of capture probes and the simplicity of the approach (2, 3). In-solution hybridization capture uses biotinylated DNA or RNA molecules (baits) to capture target regions (1–3, 9). A molar excess of biotinylated baits is hybridized with NGS libraries in solution. The resulting library-bait heteroduplexes are captured on streptavidin-coated magnetic beads. Unbound non-target molecules are washed away, and target molecules are recovered for sequencing (9, 10).

Current bait synthesis methods require large-scale oligonucleotide chemical synthesis and/or in vitro transcription. Both RNA and DNA bait generation requires synthesizing template oligonucleotides using phosphoramidite chemistry. Microarray-based synthesis generates oligonucleotides in femtomole scales with chemical coupling error rates of 10-2 - 10-3 (11, 12). Templates synthesized at small-scale require enzymatic amplification before use in hybridization capture. For RNA baits, PCR amplified oligo templates are transcribed in vitro into biotinylated RNA baits as initially described by Gnrike et al (9). However, in vitro transcription using T7 RNA polymerase can lead to amplification biases based on the templates' sequence, length, and GC content (13, 14). For DNA baits, either a small-scale template pool is enzymatically amplified (Twist Biosciences product sheet) or each bait is individually manufactured at scale (IDT product sheet).

We present a cost-effective, large-scale DNA bait synthesis method that we call Circular Nucleic acid Enrichment Reagent, or CNER (pronounced as snare). The CNER method involves circularization of target template oligos that contain a linker region to promote circularization via splint-ligation and a rare-cutter restriction enzyme site for subsequent discretization of the capture probes. Circularized templates are isothermally amplified by rolling circle amplification (RCA) with the inclusion of biotinylated nucleotides. The long RCA products are discretized into single biotinylated baits by restriction digestion (Figure 1.1). The resulting biotinylated CNER probes can be generated in microgram quantities and used for capture enrichments on streptavidin-coated beads.

Here, we demonstrate the use of the CNER method for targeted genotyping by producing a set of CNER probes to capture 23,771 SNPs in the horse genome. We use these CNERs to capture target SNPs from ten ancient horse DNA libraries of varying endogenous DNA content and DNA degradation levels. We show that the CNERs effectively perform target enrichment even in highly degraded ancient samples comparably to or better than commercially made baits and at a fraction of the cost.

Materials and Methods

DNA isolation

We selected ten ancient horse samples of varying DNA preservation (details in Supplementary Table S1 and in (15)) to test the performance of the CNER method. The samples date to the Late Pleistocene between 10,000 and 50,000 years ago, based on stratigraphic information and directly radiocarbon dated collagen (Supplementary Table S1 and in (15)). We extracted ancient DNA following (16) in a dedicated ancient DNA laboratory at the UC Santa Cruz Paleogenomics Laboratory (PGL) and following standard protocols for handling ancient DNA (17).

We isolated DNA from four modern domestic horses for capture optimization using blood samples drawn in May/June 2017 during routine veterinary checks. We used the DNeasy Blood & Tissue kit (Qiagen) following the manufacturer’s protocol.

Sequencing library preparation

We prepared NGS libraries from each horse extract using the Santa Cruz Reaction (SCR) (18). For the modern horse, we fragmented genomic DNA using 0.02U DNase I (Thermo Fisher) at 15°C for 15 min with MgCl2 before proceeding with the SCR. We prepared ancient horse DNA libraries in the dedicated clean at the PGL. For both ancient and modern samples, we divided adapter-ligated DNA into three aliquots before PCR amplification. We PCR-amplified ancient DNA libraries with Illumina unique dual index primers (19) using 2x AmpliTaq Gold 360 master mix (Thermo Fisher) at 95°C for 10 min, followed by 10-15 cycles of 95°C for 30 s, 60°C for 30 s, 72°C for 1 min, with a final extension at 72°C for 7 min followed by a hold at 12°C. We PCR amplified the modern horse libraries with Illumina unique dual index primers using 2x KAPA HiFi master mix (Roche) at 98°C for 3 min, followed by 13 cycles of 98°C for 30 s, 65°C for 20 s, 72°C for 20 s, with a final extension at 72°C for 3 min then hold at 12°C. We purified the amplified libraries with SPRI (20) beads at 0.8x ratio for the modern horse and at 1.2x for the ancient horses, quantified the DNA using Qubit 1x HS assay (Thermo Fisher), and determined library size by Fragment Analyzer (Agilent).

Horse SNP panel design

We designed the horse SNP panel for target enrichment of known nuclear SNPs based on the SNP ascertainment scheme described in (15). Briefly, we genotyped Batagai (21), CGG10022 (22), YG188.42/YT03-40 and YG303.325 (both from ref (15)) ancient horse genomes mapped to EquCab2 (GenBank: GCA_000002305.1; (23)) as described in (15), using samtools v.1.7 utilities mpileup and bcftools (24), AntCaller v1.1 (25), and GATK HaplotypeCaller 3.7 (26). We intersected variant calls from all three programs using VCFtools v0.1.16 vcf-isec (27). In downstream analyses, we used only variants called by all three programs. We also removed variants with <20 base call quality, <5X read coverage, location within 5 bp of indels, singletons and homozygous alternative alleles in all four ancient horse genomes. We selected SNPs located outside of gene boundaries and repetitive regions using the filtering strategy described in (15).

We selected the final set of 26,944 candidate variant loci for bait designing by Arbor Biosciences. Arbor provided us a list of 74,385 candidate baits. We filtered these to limit to 60K baits based on the chosen synthesis tier. We chose baits with 20 - 80% GC content, filtered out baits containing repeats using RepeatMasker and baits with strong secondary structures (ΔG > -9 kcal/mol). After filtering, we chose a final list of baits to target 22,619 variant loci to proceed with Arbor myBaits generation. The final Arbor panel targeted 2583 SNPs using one bait, 3391 SNPs using two baits, and 16645 SNPs using three baits, and 228 Y-chromosome targets representing sequence-tagged sites (STS), AMLEY, and SRY genes. All 59,528 Arbor myBaits were 80 nt long RNA probes.

For CNERs generation, we targeted the same randomly selected 22,619 autosomal SNPs, each with one 80-bp long CNERs centered at the SNP site, plus the same 228 Y chromosome targets. To test the effect of CNERs length on coverage, we selected two additional sets of 576 SNPs and designed 50bp and 100bp CNERs with SNPs at the center. In total, the horse SNP panel targets 23,771 SNPs using a total of 23,999 probes.

Horse SNP panel CNERs generation

We generated CNERs for the horse SNP panel as schematically described in Figure 1.1. We appended six deoxy-T (dT) bases at the 5’ end, and AscI restriction site and (dT)6 at the 3’ end to all horse target regions to make CNERs templates. We synthesized the templates as an DNA oligo pool using silicon chip based phosphoramidite chemistry (Twist Biosciences). We circularized 100 or 300 femtomoles of the oligo pool in a 20 µl splint ligation reaction containing 2000U T4 DNA ligase (NEB), 10U T4 PNK (NEB) and 1000fmol (dA)12 splint oligo in 1X T4 DNA ligase buffer at 37°C for 1 h followed by 25°C for 3 h and denatured at 95°C for 3 min. We amplified the circularized oligo pool in a 50 µl RCA reaction containing 30U of Phi29 polymerase (NEB), 25 pmol each of forward (5’ - AAAAAAAAAGGCGCGCC - 3’) and reverse (5’ - GGCGCGCCTTTTTTTTT - 3’) RCA primers, 2 nmol each of biotin-11-dATP (Perkin Elmer) and biotin-11-dUTP (Thermo Fisher), 25 nmol each dNTPs in 1X Phi29 buffer with BSA. After 40 - 48 h of RCA reaction at 30°C, we purified RCA products using SPRI beads (1.2x ratio) and digested with 100U AscI (NEB) for 5 h at 37°C to produce monomeric CNERs. We estimated size and concentration of RCA products before and after AscI digestion using capillary electrophoresis in a Fragment Analyzer (Agilent) with the genomic DNA kit. We purified post-digestion products using SPRI beads (2x ratio) and quantified the DNA using a Qubit (Thermo Fisher).

CNERs hybridization capture optimization

We optimized CNERs capture for adapter blocker concentration, CNER amount per reaction, and hybridization buffer compositions. To optimize adapter blocker concentration, we titrated oligonucleotide blockers at 5x - 200x molar excess to 100 - 300 ng (1.0 - 2.3 pmoles) of the modern horse libraries, 25 ng horse SNP panel

CNERs, 2.5 µg of Human c0t DNA, and 25 µg of salmon sperm DNA in 25 µl reaction, and then denatured at 95°C for 10 min. We added this DNA mixture to 25 µl prewarmed Hyb buffer (final concentrations: 6X SSPE, 6X Denhardt's Solution, 10mM EDTA, pH 8.0, 0.2% SDS) and hybridized the mixture overnight in 50 µl total reaction volume at 65°C. To optimize CNERs amount titrations, we hybridized 300ng of libraries with 30 – 90 ng of horse SNP panel CNERs and 200x molar excess oligo blockers in the Hyb buffer at 65°C overnight. We tested four hybridization buffers (HB1: 100mM MES pH 6.5 and 1M NaCl; HB2: 6X SSC, pH 7.0; HB3: 6X SSPE, pH 7.4; and HB4: 100mM Tris pH 8.0 and 1M NaCl) to capture 250ng of libraries using 50 ng CNERs overnight at 65°C. All four buffers also contained 0.1% SDS, 10mM EDTA and 10% DMSO at final concentration. We captured CNER hybridized libraries onto 30 µl MyOne C1 streptavidin beads (Thermo Fisher) at 65°C for 30 min. We washed beads three times in high stringency wash buffer (0.2X SSC, 0.1% SDS, and 10% DMSO) for 5 min each at 65°C and then three times in low stringency buffer (2X SSC and 0.1% SDS) at room temperature. We washed beads in 10mM Tris pH 8.0 before resuspending in the PCR reaction. We amplified post-captured libraries using 2x KAPA HiFi master mix (Roche) and Illumina universal amplification primers at 98°C for 3 min, followed by 15 cycles of 98°C for 30 s, 60°C for 30 s, 72°C for 30 s, with a final extension at 72°C for 5 min then hold at 12°C. We purified post-capture libraries with 0.9x SPRI beads, quantified using a Qubit (Thermo Fisher), pooled, and sequenced on an Illumina NextSeq using PE 2x150 kit.

Ancient Horse DNA capture and sequencing

For the ancient horse samples, we captured 5 µl (constant library volume with varying library mass; see Supplementary Table S2 for details) of individual ancient horse libraries using Arbor myBaits and CNERs. For both Abor myBaits and CNERs captures, we performed two experiments. In experiments A1 (CNERs) and A2 (Arbor myBaits), we followed the Arbor myBaits protocol and used 50% of capture beads for post-capture amplification and purified libraries with 1.7x SPRI as per the protocol. In experiments B1 (CNERs) and B2 (Arbor myBaits), we followed the optimized CNERs protocol, and used 100% of capture beads for PCR and 0.9x SPRI for cleanup. Finally, we performed a separate CNERs Experiment C, in which we captured libraries in 3-plex pools. In experiment C, we also used 100% of captured beads for PCR amplification and purified the post-capture libraries with 0.9x SPRI.

For all experiments using CNERs, we used 2 µl (~40ng) of the horse SNP panel CNERs. For a single sample, UAM:ES:27502, for which little material remained at the start of the experiment, we used only 2 µl of library CNERs in both experiment A and B. For all other samples, we used 5 µl libraries for captures. We added 200x adapter blocking oligos, 2.5 µg of Human c0t DNA and 25 µg of salmon sperm DNA to these library-CNERs to a total of 30 µl volume, and then denatured at 95°C for 10 min. We preincubated 30 µl of HB4 at 62°C for 5min, mixed with denatured library/CNERs/blockers mixture and hybridized at 62°C for 19.5 hr. We enriched post-hybridization libraries onto streptavidin beads as in the optimization experiments except both low and high stringency wash steps were done at 65°C.

For CNERs experiment C (pooled capture), we hybridized 67 - 100 ng of libraries for each of three samples with similar endogenous content with 40 - 60 ng CNERs (Supplementary Table S2). We repeated the individual capture for UAM:ES:26433, rather than including it in a pool, as it had the lowest pre-capture endogenous content. We did not perform pooled captures for Arbor myBaits as it was not recommended by the manufacturer.

For all captures using Arbor myBaits, we used 5 µl of the same ancient horse libraries that we used in CNERs captures. We used unopened vial of the Arbor myBaits Horse SNP panel. Although the baits had been stored at -80°C continuously since production, they were 15 months older than the labeled use-by date. We followed Arbor Biosciences capture protocol v3 with recommended modifications of hybridization at 55°C for 41 h for ancient DNA.

We used different approaches to post-capture library amplification in experiments A compared to experiments B. In A, we resuspended capture beads in 30 µl 10mM Tris pH 8.0 buffer. We then used 15 µl of the resuspended beads in 20 cycles of PCR amplification with 2x KAPA HiFi. We then purified the product with 1.7x SPRI, as recommended by Arbor. For B, we resuspended capture beads in 20 µl 10mM Tris pH 8.0 buffer and used all of it in a 50 µl PCR reaction and performed 20 cycles of amplification, followed by purification with 0.9x SPRI.

All post-capture libraries were Qubit (Thermo Fisher) quantified, pooled, and sequenced on an Illumina NextSeq with a PE 2x75 kit.

Bioinformatic processing

We trimmed adapter sequences from the reads and merged overlapping paired end reads using SEQPREP2 (https://github.com/jeizenga/SeqPrep2). We mapped merged and unmerged reads to the EquCab2 reference (23) genome using BWA ALN - v0.7.17-r1188 (28). We marked and removed duplicated reads using PICARD MARKDUPLICATES - v2.21.7 and calculated capture metrics using PICARD COLLECTHSMETRICS - 2.21.7 (http://broadinstitute.github.io/picard). We determined read coverage at target SNPs using BEDTOOLS MULTICOV - v2.29.1. We plotted SNP coverage against CNERs length, GC content, and percent targets using custom python scripts (https://github.com/bsun210/CNERs_ancient_horses). We used BEDTOOLS INTERSECT - v2.29.1 to find sequence reads mapping to the target SNPs to calculate the position of SNPs relative to the sequence read insert size. We determined genotype likelihoods for the ancient horses using ANGSD - v0.935-52-g39eada3 with -GL 2 -minMapQ 20 -nThreads 24 -doGlf 2 -doMajorMinor 1 -SNP_pval 1e-6 -doMaf 1 options (29). We analyzed population clustering and ancestry using PCANGSD - v1.10 with default settings (30). We used PRCOMP and FACTOEXTRA R packages (31) for principal component analysis (PCA). We calculated endogenous content (proportion of unique reads aligned to the horse genome), library complexity (proportion of uniquely mapped non-duplicated molecules) and insert size distribution using the pipeline described in (15).

We assessed whether the SNP coverage for CNERs with different lengths, changes in endogenous content, library complexity, and insert size between pre and post-capture libraries are normally distributed using the Shapiro-Wilk test. All these groups are not normally distributed; hence we performed a nonparametric Mann-Whitney Wilcoxon (MWW) rank test for comparison between groups. For comparison of normalized coverage distribution across GC bins for various experimental groups, we used two sample Kolmogorov-Smirnov (KS) tests for goodness of fit.

Results

The CNER method is designed to generate large amounts of biotinylated baits for hybridization capture (Figure 1.1). CNER templates are synthesized as oligonucleotides with oligo-dT linkers at both 5' and 3' ends to facilitate circularization using a complementary, oligo-dA splint. Because the linkers are oligo-dT, this design limits the impact of incomplete oligonucleotide chemical synthesis errors at the template ends. In the 3' end upstream of the oligo-dT, a rare-cutter restriction enzyme recognition site (RES) is also incorporated (Figure 1.1). Oligo-dT and rare cutter RES are appended to all target sequences such that all CNER templates have uniform ends to facilitate bulk circularization by splint ligation using an oligo-dA splint adapter (Figure 1).

After circularization, CNER templates are bulk amplified by rolling circle amplification (RCA) using high processivity phi29 DNA polymerase. The RCA reaction includes biotin-dATP and biotin-dUTP (an inexpensive and widely available alternative for biotinylated dTTP) in the reaction to generate biotinylated products.

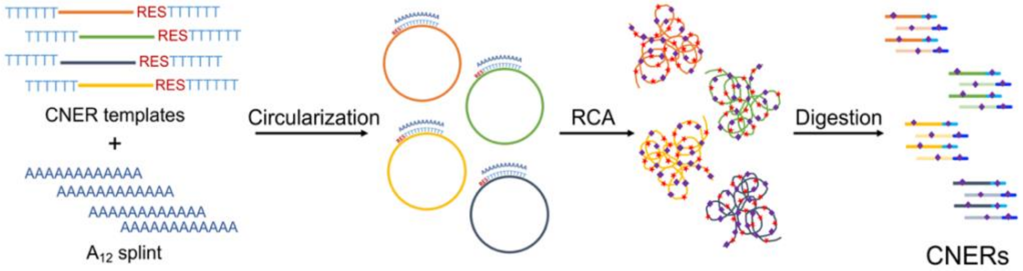


Figure 1.1 Circular Nucleic acid Enrichment Reagent method.

An oligonucleotide template pool containing restriction enzyme recognition sites (RES) and oligo-dT linkers is circularized by an oligo-dA splint adapter mediated ligation.

Circularized templates are isothermally amplified using oligo-dA and oligo-dT oligos by rolling circle amplification (RCA). RCA products are then digested with restriction enzymes to generate CNERs. CNERs generate both strands (dark and light shades of colors) of the templates. Biotinylated nucleotides (purple diamonds) are incorporated during amplification.

An oligo-dA forward primer and oligo-dT reverse primer initiate forward and reverse RCA reactions. Thus, the RCA products for each CNER template is double-stranded, regardless of which strand the original CNER template was designed against (Figure 1.1). Further, inclusion of both forward and reverse primers facilitate branched amplification during RCA to increase yield. The RCA makes many of copies of the CNERs as concatemers, a single restriction enzyme digestion of which produces monomeric, biotinylated capture probes (Figure 1.1). The monomeric CNERs can therefore be used as baits to capture and enrich target molecules on streptavidin-coated beads for sequencing.

We designed a horse SNP panel with 23,771 randomly selected SNPs from a list of high confidence variant sites ascertained in four ancient horse genomes (15). Chemical synthesis of oligo templates for this panel yielded a 215 ng (6.3 picomoles) pool. RCA amplification of 100 femtomoles (~3.3 ng) bulk circularized template pool generated 611 ng of double-stranded high-molecular weight DNA (~77 kB average size, Supplementary Figure S1A), restriction digestion of which generated 499 ng of monomeric CNERs with 114 bp average size (Supplementary Figure S1B). The presence of double-stranded DNA indicates that the CNERs method generates probes against both strands of the target region. In a separate experiment, we increased the input template to 300 femtomoles. The protocol yielded 1.57 µg CNERs in that experiment. Thus, we estimate 100 femtomoles (~3.3 ng) of circularized CNER templates produces ~500 ng of CNERs using the protocol as described.

CNER hybridization optimization

We optimized in-solution hybridization conditions for the horse SNP panel CNERs using the modern horse DNA libraries (see Supplementary Data). We tested hybridization capture reactions with increasing amounts of adapter blocking oligos to prevent cross-hybridization of library molecules (32) with a constant amount of CNERs. In a separate set of experiments, we tested increasing amount of CNERs with a constant amount of blocking oligos. Both increasing amount of blocking oligos and CNERs modestly improved the enrichment efficiency (Supplementary Table S3, Supplementary Figure S2A and S2B). We note that conventional hybridization buffer like those used by Arbor myBaits for RNA baits (33) might be suboptimal for DNA baits. Therefore, we tested four hybridization buffers (HB) to improve the enrichment efficiency for CNERs. Captures in HB4 produced >50% (by Picard metrics) bases on or near targets for the modern horse libraries (Figure 1.2A). Additives used in conventional hybridization buffers like Denhardt's solution and trimethyl ammonium chloride did not improve and or lowered the percentage of on or near target bases (Supplementary Figure S2C). Hybridization at 62°C and 65°C also resulted in similar enrichment efficiency (Supplementary Figure S2D).

Existing capture bait synthesis methods use different probe lengths and tiling to optimize for the GC content of target regions (34, 35). We designed CNERs with three different lengths to test the effect of CNER length on SNP coverage. The 80bp CNERs produce higher SNP coverage than either 50bp or 100bp CNERs (Figure 1.2B) consistently across various hybridization conditions (Supplementary Figure S3). Further, target regions within 43% - 65% GC bins, which are 47% of the total target SNP regions (average GC = 43.8%), consistently resulted in >1 normalized coverage (Figure 1.2C, Supplementary Figure S4).

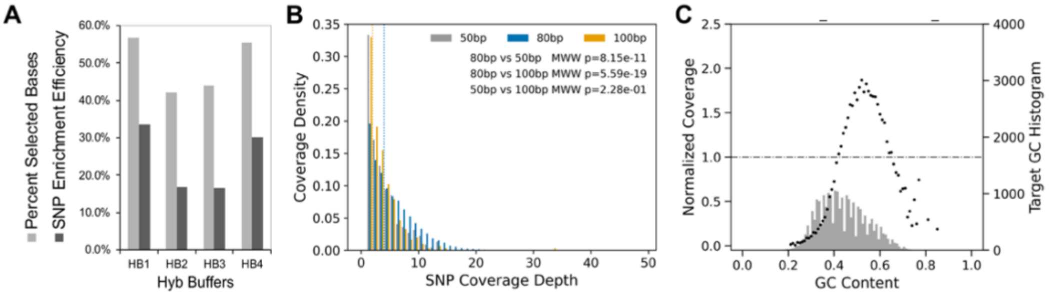


Figure 1.2 Optimization of CNERs hybridization capture of SNPs in four modern horse samples.

(A) *Enrichment efficiency for four hybridization buffers with pH varying from 6.5 - 8.0 (HB1 - 4). Light grey bars show the Percent Selected Bases determined using Picard tools and dark grey bars show the SNP enrichment efficiency. Values presented are the average of three experiments for HB1 and HB4 buffers and exact values for a single experiment for HB2 and HB3.* ***(B)*** *Histogram density plots of SNP coverage depth for three CNER lengths. SNPs captured with 80bp CNERs (blue bars) result in significantly higher coverage compared to SNPs captured with 50bp (grey bars) or 100bp (orange bars) CNERs; p-value is from a Mann-Whitney Wilcoxon test. Dotted lines indicate the mean coverage for each CNERs length.* ***(C)*** *Mean of normalized coverage (primary Y-axis) plotted across GC content of CNER target regions show that regions with 43% - 65% GC have sample-normalized coverage of 1 or higher. A histogram of GC bins across the target regions is shown in the secondary Y-axis.*

CNERs efficiently capture ancient DNA target SNPs

We extracted DNA from ten horse bones collected from Late Pleistocene age permafrost deposits in Alaska, USA and Chukotka, Russia (Supplementary Table S1 and (15)). Sequence reads generated from each of these samples, mapped to the EquCab2 reference genome, provided estimates of endogenous DNA content. Before SNP enrichment, the ancient horse DNA libraries had 18.4% median reads mapped to the horse genome, across a wide range (6.0% - 91.2%, ‘preCap’ in Figure 3A, Supplementary Table S2).

SNP enrichments using both DNA based CNERs and RNA based Arbor myBaits increased the proportion of reads in the sequencing library that mapped to the reference genome, indicating successful target enrichment. Enrichment using CNERs improved median precent of mapped reads to 37.9% in experiment A (individual captures following the Arbor myBaits protocol), and 30.5% in experiment B (individual captures following the CNERs protocol), and 40.1% in experiment C (pooled-captures with CNERs protocol). Arbor myBaits resulted in 28.8% in experiment A (individual capture following the Arbor myBaits protocol), and 21.1% in experiment B (individual capture following the CNERs protocol) (Figure 1.3A, Supplementary Table S4. Comparison of CNERs experiments B versus C show a consistent proportion of mapped reads when a sample was captured individually versus as part of a pool (Figure 1.3A). The differences between capture probes and protocols are not significant by Mann-Whitney Wilcoxon test.

Different SPRI bead ratio used in the post-capture purification steps did not affect the proportion of mapped reads (Figure 1.3A). However, the different SPRI ratio resulted in different proportions of merged and unmerged reads identified during data analyses. Short insert size of aDNA molecules result in overlapping read pairs which are merged during data processing, hence called as merged reads. Read pairs that did not overlap are processed as unmerged read pairs. Following the Arbor myBaits protocol which uses 1.2x SPRI beads ratio (experiments A) resulted in a higher proportion of merged compared to unmerged reads for both Arbor myBaits and CNERs (Supplementary Figure S5A, Supplementary Table S4). All experiments that followed the CNERs cleanup protocol resulted in equal proportions of merged and unmerged reads regardless of probes, due to the lower SPRI beads ratio (0.9x) used during the post-amplification cleanup. Across all experiments, a greater proportion of merged reads mapped to the reference genome compared to unmerged reads, as expected for aDNA (Supplementary Figure S5B).

Previous studies used Picard's program CollectHsMetric to measure the success of target enrichment (36). This tool reports coverage of the targeted base and 100bp flanking regions when determining 'Percent Selected Bases'. We used this metric during the optimization experiments to compare the performance of CNERs to current standards. However, this metric overestimates the SNP enrichment success by including the regions around the target SNP site. Therefore, we elected to measure the success of SNP enrichment in ancient horses by defining 'SNP enrichment efficiency' as the percentage of all or mapped reads that are exactly mapped to the target SNPs. This is a straightforward but more practically important measure of SNP enrichment success. For the modern horse captures with CNERs, hybridization in HB4 at 65°C for 18 - 20 hour produced ~30% SNP enrichment efficiency for mapped reads (Figure 1.2A). We followed these hybridization conditions to capture ancient horse samples.

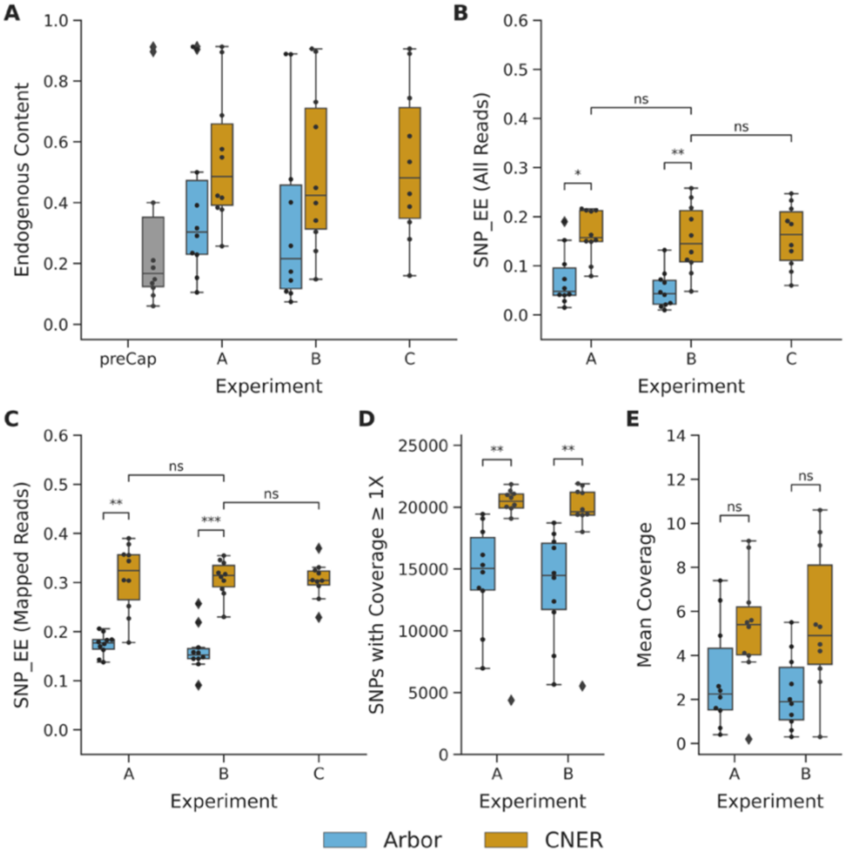


Figure 1.3 SNP capture with CNERs and Arbor myBaits for ancient horse samples. ***(A)*** *Endogenous content measured as proportion of reads mapping to horse reference genome for ten ancient horse samples before capture enrichment (grey bars), proportion of mapped reads after capture with Arbor myBaits (cyan), and proportion of mapped reads after capture with CNERs (yellow). SNP enrichment efficiency measured as proportion of total reads* ***(B)*** *and mapped reads* ***(C)*** *covering the target SNPs for CNERs and Arbor myBaits.* ***(D)*** *Number of target SNPs covered by at least one read.* ***(E)*** *Mean coverage of target SNPs at one million raw read pairs. Mann-Whitney Wilcoxon test p values are indicated as ns (5.00e-02 < p <= 1.00e+00), * (1.00e-02 < p <= 5.00e-02), ** (1.00e-03 < p <= 1.00e-02) and *** (1.00e-04 < p <= 1.00e-03).*

SNP enrichment efficiency, or the proportion of reads mapping to the target SNPs, was significantly higher when using CNERs compared to when using Arbor myBaits. In experiments A (Arbor myBaits protocol), the median SNP enrichment efficiency was 15.7% for CNERs vs 4.8% for Arbor (MWW $p < 0.05$). In experiments B (CNERs protocol), the median SNP enrichment efficiency was 14.5% for CNERs vs 4.3% for Arbor myBaits (MWW $p < 1e\text{-}2$; Figure 1.3B, Supplementary Table S4). This pattern holds when considering only reads that map to the reference genome. Experiments A (Arbor myBaits protocol) resulted in median enrichment efficiencies of mapped reads of 32.4% for CNERs vs 17.7% for Arbor myBaits (MWW $p < 1e\text{-}2$), and experiments B (CNERs protocol) resulted in median efficiencies of mapped reads of 31.5% for CNERs vs 15.2% for Arbor myBaits (MWW $p < 1e\text{-}3$; Figure 1.3C). The pattern is also consistent when considering merged and unmerged reads separately, both for all reads and mapped reads (Supplementary Figure S5C and S5D), although unmerged reads always had significantly lower enrichment efficiency compared to merged reads (Supplementary Figure S5D, Supplementary Table S4). Finally, the enrichment efficiency when using CNERs was consistent between individually captured libraries and captures performed in pools (Figures 1.3B and C).

To test the potential impact of differences in sequencing depth, we subsampled data to one million read pairs per sample in experiments A and B. For this analysis, we considered only the 22,619 target SNPs that were common between CNERs and Arbor myBaits. For experiments A (Arbor myBaits protocol), this read depth resulted in a median of 90.5% (20,479) of target SNPs covered by at least one unique read using CNERs versus 66.5% (15,038) for Arbor myBaits (MWW $p < 1e\text{-}2$; Figure 3D, Supplementary Table S4). We observed a similar trend when following the CNERs protocol (experiments B; Figure 1.3D). At this coverage, CNERs captures have fewer SNP dropouts compared to Arbor myBaits captures, as estimated using cumulative distribution plots of SNP coverage as percentage of SNPs less than the x-fold mean coverage (Supplementary Figure S6). When averaged across the 10 horse data sets at this standard coverage, CNERs captures resulted in 2.5-fold higher average SNP coverage than Arbor myBaits (an average of 5.4 reads per SNP compared to an average of 2.2 reads per SNP when using the Arbor myBaits protocol (experiments A; Figure 1.3E), and an average of 4.9 reads per SNP compared to an average of 1.9 reads per SNP when following the CNERs protocol (experiments B; Figure 1.3E). The average coverage was not significantly different by MWW test due to one outlier sample (UAM:ES:27502), which was the sample for which we had to reduce library volume going into CNERs captures and has low SNP coverage.

We evaluated target coverage uniformity using fold-80 base penalty, which estimates additional sequencing required to bring 80% of the zero-coverage targets to mean coverage depth. The smaller the fold-80 base penalty, the more uniform the coverage is across all target regions (37). The average fold-80 base penalty is 3.7 for CNERs and 5.3 for Arbor myBaits, suggesting that CNERs produces more uniform coverage across all target SNPs.

We explored whether probe length or GC content explained coverage unevenness among the ancient horses. As observed in the modern horse enrichments, enrichment of ancient horses resulted significantly higher SNP coverage for CNERs targeting 80bp regions compared to 50bp or 100bp (Supplementary Figure S7). The statistical degree of significance of these comparisons as estimated from MWW test p-values (Supplementary Figure S7) differed among the ancient horses due to differences in percent mapped reads. Enrichments using CNERs resulted in higher normalized coverage for SNPs in target regions that had 42-66% (mode ~55%) GC content compared to SNP targets in other GC contents and to Arbor myBaits capture data in this GC bin (Supplementary Figure S8). Arbor myBaits resulted in higher SNP normalized coverage for target regions with 30-45% GC content (mode ~37% GC) compared to other GC contents and to CNERs capture data in this GC bin. While this indicates a shift towards lower GC preference for Arbor myBaits and higher GC preference for CNERs, the difference in coverage across GC bins is not statistically different by KS test (Supplementary Figure S8).

We next compared CNERs captures and Arbor myBaits captures in the mean normalized coverage at 100 bp upstream and downstream regions of target SNPs to assess whether coverage around the SNP target region influenced coverage unevenness. We designed only one CNER per target SNP, centered in the target region, resulting in maximum coverage depth for SNPs and reduced coverage for the surrounding region (Supplementary Figure S9). Arbor myBaits designed up to three baits per target SNP, tiled 20 bp from 5' end, which resulted in an expected maximum coverage for ~20bp region to the right of the target SNP (Supplementary Figure S9). These differences in coverage profile between CNERs and Arbor myBaits are significant by KS test. Post-capture purification steps did not affect the coverage around SNPs; both experiments A (Arbor myBaits protocol) and experiments B (CNERs protocol) resulted in similar coverage profiles when comparing enrichments using same probes (Supplementary Figure S9).

CNERs and Arbor myBaits produce similar genotypes

We calculated genotype likelihoods for target SNPs using the capture data. We did not include sample UAM:ES:27502 because it had few genotyped sites. Average concordance of genotypes of nine ancient horses between experiment A (Arbor myBaits protocol) and B (CNERs protocol) is 97.9% for Arbor myBaits data and 98.1% for CNERs data (Supplementary Figure S10, Supplementary Table S5). To increase the read depth for individual SNPs, we merged bam files from the two experiments and called genotypes on the merged data. With merged data, both CNERs and Arbor myBaits genotyped between 4,394 – 13,330 sites with 96.7% - 99.5% concordance for individual horses (Figure 1.4A). On average, genotypes called on Arbor myBaits and CNERs data concur 98.6%.

CNERs and Arbor myBaits captured reads with different base substitution patterns in the target SNPs (Figure 1.4B). Of the total 18,994 genotyped sites among the nine ancient horses, 13,893 sites were captured using both probes, 1,334 sites were only captured by Arbor myBaits and 3,767 sites were only captured by CNERs data. CNERs capture more GC transversions compared to Arbor myBaits (Figure 1.4B) because they more efficiently capture higher GC regions (Supplementary Figure S8). While CNERs and Arbor myBaits capture reads with comparable patterns of cytosine deamination at the ends of reads (Supplementary Figure S11), Arbor myBaits captured more SNPs with transition substitutions (11.5% vs 4.5% for CNERs vs 0.4% shared in both probes, Figure 1.4B). This pattern may arise because the right shifted tiling design preferentially enriches for SNPs at the ends of aDNA molecules (Supplementary Figure S12) where transition substitutions occur due to cytosine deamination. Alternatively, CNERs enrich for aDNA fragments with SNPs at the center of the read (Supplementary Figure S12), which may lead to higher coverage at SNP sites compared to Arbor myBaits (Supplementary Figure S9).

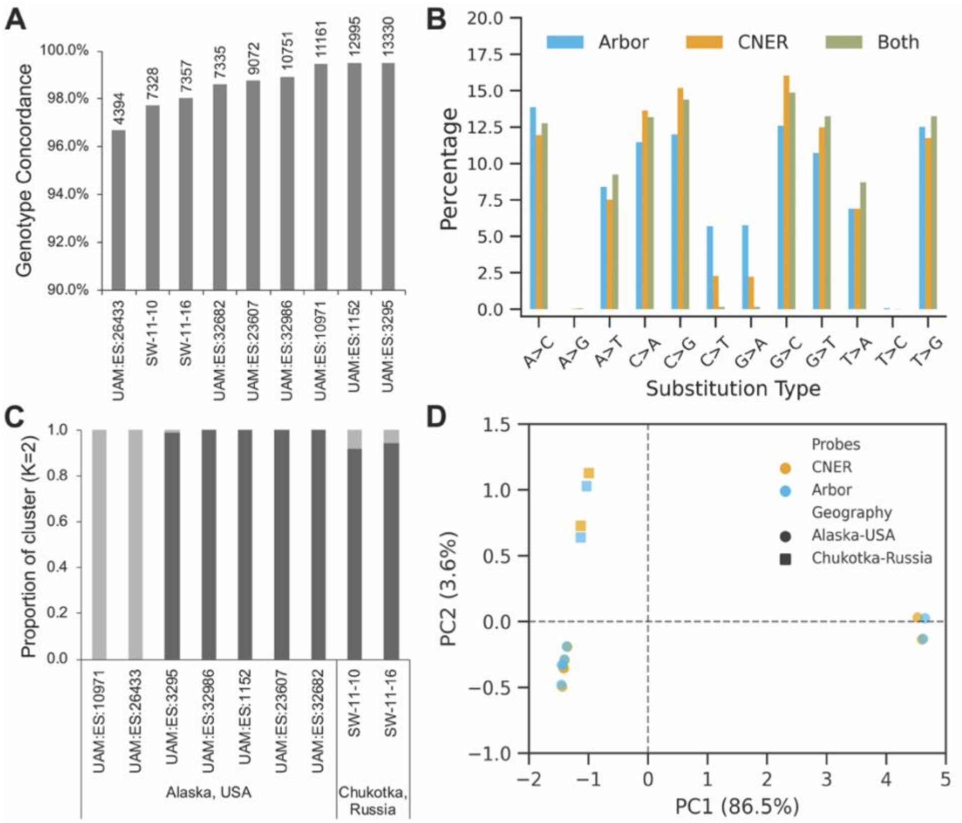


Figure 1.4 Genotyping and estimated evolutionary relationships between the ancient horse samples.

(A) *Genotype concordance between SNP capture data generated using CNERs and Arbor myBaits. Numbers above the bars indicate the sites genotyped by both methods in a given horse sample.* ***(B)*** *Percentage of substitution types shared between (green) and unique to CNERs (yellow) and Arbor myBaits (cyan).* ***(C)*** *Admixture analysis with K=2 separated the ancient horses into two lineages regardless of their geographic location.* ***(D)*** *Principal component analysis of genotype likelihood covariance matrix of 23,771 nuclear SNP sites in nine ancient horses. Transitions are filtered out for* *population analyses due to cytosine deamination in aDNA. PC1 segregated horses into two major clades and PC2 separated horses into the Western (Chukotka) and Eastern (Alaska) Beringian populations.*

We used the enriched genotypes to explore the evolutionary relationships between the nine ancient horses for which we generated data. Admixture analysis identified two main ancestry components, both for data generated using CNERs (Figure 1.4C) and Arbor myBaits captures (Supplementary Figure S13). Principal component (PC) analysis of genotype likelihood covariance also segregated ancient horses into two major clusters (Figure 1.4D), with similar patterns observed when using CNERs or Arbor myBaits data. The first principal component (PC1) roughly corresponds to ancestry as in Figure 1.4C, and PC2 reflects geographic origin either in Chukotka, Russia (Western Beringia) or Alaska, USA (Eastern Beringia). This pattern is consistent among probe types and with horse population structure previously inferred from whole-genome and mitochondrial data (15).

Discussion

Targeted sequencing can provide a cost-effective method for data generation for many comparative genomics applications, in particular when the samples of interest contain only trace amounts of degraded DNA. However, the high cost of producing hybridization baits hinders the widespread adoption of this approach. Our approach, which we call Circular Nucleic acid Enrichment Reagent method, reduces both the cost and time required for generation of microgram quantities of probes. Incorporation of poly-dT overhangs at both ends in the CNER template design overcomes end synthesis errors in long oligonucleotide baits. The length of the poly-dT limits the circularization of templates by splint ligation using the poly-dA oligo. Poly-dA mediated splint ligation ensures that only templates with a certain length of poly-dT are amplified by RCA, thus eliminating incompletely synthesized baits. These template design features and isothermal amplification using RCA overcome many of the artifacts induced by PCR amplification of template oligo pools like non-specific amplification and generation of heterogenous products (Twist Bioscience's technical note). Further, standard PCR amplification requires inclusion of specific primer binding sequences at the ends that increase oligo length (9) and may interfere with hybridization capture. Future comparison of the CNERs methods with other PCR-based oligonucleotide amplification methods would be useful to explore the role of amplification biases in hybridization efficiency.

We optimized the hybridization conditions for the CNERs which differed from conventional hybridization conditions used for RNA baits. Enrichments using CNERs reduces the hybridization time to overnight incubation (18 - 20 hours) instead of the 48 - 72 hours required in conventional capture methods for degraded DNA (33, 35). This increase in efficiency may be useful in clinical diagnostics. Further, conventional baits are designed with multifold tiling baits per target (34, 35) to achieve uniform coverage across different GC regions, but still underperform for target regions with >50% GC content (34, 36). We designed only one CNER tiling per SNP target region to save both CNERs production cost and sequencing cost. CNERs capture results in higher coverage for target regions with 45 - 75% GC content than regions with other GC contents, similar to other DNA baits (36), whereas Arbor myBaits produced higher coverage for regions with 30-45% GC, similar to other RNA baits (35, 36). Difference in the AT/GC bonding strength might differently influence the melting temperature of DNA-RNA heteroduplex and double stranded DNA molecules, which could lead to the observed coverage differences between the DNA and RNA baits for target regions with different GC content. It would be interesting to test whether multi-tiling CNERs for target regions with lower GC content brings their coverage closer to the sample mean coverage. Multi-tiling and probe length also increase the coverage for regions around the targeted region (33, 34). This might be desired for some applications like exome capture, but it will reduce the cost-effectiveness of genotyping-by-sequencing (GBS). CNERs achieve highest coverage at the target SNP sites compared to adjacent regions which is desired for GBS applications.

To demonstrate the utility of the CNERs approach for GBS, we genotyped ~23k nuclear SNPs in ten ancient horses using both DNA based CNERs and a commercially available RNA baits from Arbor myBaits. We found that SNP enrichment efficiency using CNERs was consistent across most of our ancient samples, despite their variability in pre-enrichment precent mapped reads (endogenous content). Further, CNERs provided two-fold higher SNP enrichment efficiency compared to Arbor myBaits. CNERs required only one probe per target SNP and enriched a greater number of targeted sites with maximal read depth at the target SNP site. Two-fold higher enrichment efficiency could be due to enrichment of both strands of target regions by the CNERs probes compared to one targeted strand by RNA baits from Arbor. This could be tested using double stranded RNA baits (36). Both admixture and PC analysis of genotype likelihoods grouped the ancient horses into two major clusters (Figure 4), like the results based on whole genomes (15). Future work using the horse SNP panel with a more geographically and temporally extensive sampling of ancient horses will provide new insights into the history of movement and gene flow among Late Pleistocene horses.

Although we focused on generating data from individual horse bones, CNERs can also be used for targeted DNA capture and sequencing from other sample types that are difficult to genotype by conventional methods (38). Cell-free and circulating tumor DNA (cf/ctDNA) isolated from liquid biopsies, for example, can be used to identify mutation burden in cancer patients, disease carrier status, and for noninvasive prenatal testing (39). DNA isolated from environmental samples like water and air and from ancient sediments can be used to reconstruct present and past environments noninvasively (40). DNA isolated from single rootless hair can be used to solve forensic cases (41). All these sample types are preserved as highly fragmented DNA, however, and often in complex mixtures, where targeted capture using CNERs provides a straightforward approach to generating useful comparative data (42).

The CNER method can be extended to generate whole genome enrichment (WGE) probes. Genome fragments of a reference or related species can be circularized by bridge adapters to included restriction enzyme sites, amplified, and digested as in oligo templates to make WGE-CNERs. These would be a DNA alternative for the whole-genome in-solution capture (WISC) method’s RNA baits (33). WGE is valuable when exploring an unknown organism or enriching a taxon in mixtures, as well as when analyzing aDNA samples with low endogenous content. WGE can also be used to generate low-coverage genomes of a few individuals for SNP ascertainment, from which a target SNP panel for population studies can be designed. We expect the CNER method adopted by futures studies for various GBS and WGE applications.

Data Availability

All raw sequencing data generated for this project are submitted to the SRA database under BioProject accession number PRJNA785663.

Supplementary Data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors thank the members of the UCSC Paleogenomics lab, especially Dr. Rachel Meyer and Remy Nguyen for their critical comments and proofreading the manuscript. We are grateful to Pamela Groves, Daniel Mann and the University of Alaska Museum-Earth Science collections for providing us access to the Alaskan ancient horse specimens analyzed in this study. We are also grateful to Love Dalen, Sergey Vartanyan, Eleftheria Palkopoulou and Alexei Tikhonov for providing the Eurasian ancient horse samples analyzed in this study. We thank Neda DeMayo, Celeste Carlisle and 'Return to Freedom' foundation for donating modern horse blood specimens.

Funding

This work was supported by the Office of Research, Industry Alliances & Technology Commercialization of the UC Santa Cruz [to REG and B Sundararaman]; the National Institute of Justice [Graduate Research Fellowship grant number 2020-R2-CX-0029 to

B Sundararaman and grant number 2020-DQ-BX-0014 to REG]; the UCSC Chancellor's Dissertation Year Fellowship [to AOV]; the National Science Foundation [grant number ARC-1417036 to B Shapiro] and the Institute of Museum and Library Services [grant number MG-30-17-0045-17 to B Shapiro].

References

1. Mamanova,L., Coffey,A.J., Scott,C.E., Kozarewa,I., Turner,E.H., Kumar,A., Howard,E., Shendure,J. and Turner,D.J. (2010) Target-enrichment strategies for next-generation sequencing. Nat. Methods, 7, 111–118.

2. Gasc,C., Peyretaillade,E. and Peyret,P. (2016) Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. Nucleic Acids Res., 44, 4504–4518.

3. Gaudin,M. and Desnues,C. (2018) Hybrid Capture-Based Next Generation Sequencing and Its Application to Human Infectious Diseases. Front. Microbiol., 9, 2924.

4. Hodges,E., Xuan,Z., Balija,V., Kramer,M., Molla,M.N., Smith,S.W., Middle,C.M., Rodesch,M.J., Albert,T.J., Hannon,G.J., et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nat. Genet., 39, 1522–1527.

5. Albert,T.J., Molla,M.N., Muzny,D.M., Nazareth,L., Wheeler,D., Song,X., Richmond,T.A., Middle,C.M., Rodesch,M.J., Packard,C.J., et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat. Methods, 4, 903–905.

6. Okou,D.T., Steinberg,K.M., Middle,C., Cutler,D.J., Albert,T.J. and Zwick,M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat. Methods, 4, 907–909.

7. Hodges,E., Rooks,M., Xuan,Z., Bhattacharjee,A., Benjamin Gordon,D., Brizuela,L., Richard McCombie,W. and Hannon,G.J. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc., 4, 960–974.

8. Burbano,H.A., Hodges,E., Green,R.E., Briggs,A.W., Krause,J., Meyer,M., Good,J.M., Maricic,T., Johnson,P.L.F., Xuan,Z., et al. (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. Science, 328, 723–725.

9. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C., et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol., 27, 182–189.

10. Maricic,T., Whitten,M. and Pääbo,S. (2010) Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE, 5, e14004.

11. Kosuri,S. and Church,G.M. (2014) Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods, 11, 499–507.

12. Song,L.-F., Deng,Z.-H., Gong,Z.-Y., Li,L.-L. and Li,B.-Z. (2021) Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data

Storage: Challenges and Opportunities. Front. Bioeng. Biotechnol., 9, 689797.

13. Duftner,N., Larkins-Ford,J., Legendre,M. and Hofmann,H.A. (2008) Efficacy of RNA amplification is dependent on sequence characteristics: implications for gene expression profiling using a cDNA microarray. Genomics, 91, 108–117.

14. Conrad,T., Plumbom,I., Alcobendas,M., Vidal,R. and Sauer,S. (2020) Maximizing transcription of nucleic acids with efficient T7 promoters. Commun. Biol., 3, 439.

15. Vershinina,A.O., Heintzman,P.D., Froese,D.G., Zazula,G., Cassatt-Johnstone,M., Dalén,L., Der Sarkissian,C., Dunn,S.G., Ermini,L., Gamba,C., et al. (2021) Ancient horse genomes reveal the timing and extent of dispersals across the Bering Land Bridge. Mol. Ecol., doi: 10.1111/mec.15977.

16. Dabney,J., Knapp,M., Glocke,I., Gansauge,M.-T., Weihmann,A., Nickel,B., Valdiosera,C., García,N., Pääbo,S., Arsuaga,J.-L., et al. (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci USA, 110, 15758–15763.

17. Fulton,T.L. and Shapiro,B. (2019) Setting up an ancient DNA laboratory. Methods Mol. Biol., 1963, 1–13.

18. Kapp,J.D., Green,R.E. and Shapiro,B. (2021) A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J. Hered., 112, 241–249.

19. Kircher,M., Sawyer,S. and Meyer,M. (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res., 40, e3.

20. Rohland,N. and Reich,D. (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res., 22, 939–946.

21. Librado,P., Der Sarkissian,C., Ermini,L., Schubert,M., Jónsson,H., Albrechtsen,A., Fumagalli,M., Yang,M.A., Gamba,C., Seguin-Orlando,A., et al. (2015) Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc Natl Acad Sci USA, 112, E6889-97.

22. Schubert,M., Jónsson,H., Chang,D., Der Sarkissian,C., Ermini,L., Ginolhac,A., Albrechtsen,A., Dupanloup,I., Foucal,A., Petersen,B., et al. (2014) Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci USA, 111, E5661-9.

23. Wade,C.M., Giulotto,E., Sigurdsson,S., Zoli,M., Gnerre,S., Imsland,F., Lear,T.L., Adelson,D.L., Bailey,E., Bellone,R.R., et al. (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. Science, 326, 865–867.

24. Li,H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27, 2987–2993.

25. Zhou,B., Wen,S., Wang,L., Jin,L., Li,H. and Zhang,H. (2017) AntCaller: an accurate variant caller incorporating ancient DNA damage. Mol. Genet. Genomics, 292, 1419–1430.

26. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., Van der Auwera,G.A., Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D., et al. (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. BioRxiv, doi: 10.1101/201178.

27. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., et al. (2011) The variant call format and VCFtools. Bioinformatics, 27, 2156–2158.

28. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760.

29. Korneliussen,T.S., Albrechtsen,A. and Nielsen,R. (2014) ANGSD: analysis of next generation sequencing data. BMC Bioinformatics, 15, 356.

30. Meisner,J. and Albrechtsen,A. (2018) Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics, 210, 719–731.

31. Kassambara,A. and Mundt,F. (2020) Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. https://cran.r-project.org/web/packages/factoextra/index.html (26 December 2022, date last accessed).

32. Lee,H., O’Connor,B.D., Merriman,B., Funari,V.A., Homer,N., Chen,Z.,

Cohn,D.H. and Nelson,S.F. (2009) Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. BMC Genomics, 10, 646.

33. Carpenter,M.L., Buenrostro,J.D., Valdiosera,C., Schroeder,H., Allentoft,M.E., Sikora,M., Rasmussen,M., Gravel,S., Guillén,S., Nekhrizov,G., et al. (2013) Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. Am. J. Hum. Genet., 93, 852–864.

34. Samorodnitsky,E., Datta,J., Jewell,B.M., Hagopian,R., Miya,J., Wing,M.R., Damodaran,S., Lippus,J.M., Reeser,J.W., Bhatt,D., et al. (2015) Comparison of custom capture for targeted next-generation DNA sequencing. J. Mol. Diagn., 17, 64–75.

35. Cruz-Dávalos,D.I., Llamas,B., Gaunitz,C., Fages,A., Gamba,C., Soubrier,J., Librado,P., Seguin-Orlando,A., Pruvost,M., Alfarhan,A.H., et al. (2017) Experimental conditions improving in-solution target enrichment for ancient DNA. Mol. Ecol. Resour., 17, 508–522.

36. Zhou,J., Zhang,M., Li,X., Wang,Z., Pan,D. and Shi,Y. (2021) Performance comparison of four types of target enrichment baits for exome DNA sequencing. Hereditas, 158, 10.

37. So,A.P., Vilborg,A., Bouhlal,Y., Koehler,R.T., Grimes,S.M., Pouliot,Y., Mendoza,D., Ziegle,J., Stein,J., Goodsaid,F., et al. (2018) A robust targeted sequencing approach for low input and variable quality DNA from clinical samples. NPJ Genom. Med., 3, 2.

38. Diaz,L.A. and Bardelli,A. (2014) Liquid biopsies: genotyping circulating tumor DNA. J. Clin. Oncol., 32, 579–586.

39. Szilágyi,M., Pös,O., Márton,É., Buglyó,G., Soltész,B., Keserű,J., Penyige,A., Szemes,T. and Nagy,B. (2020) Circulating Cell-Free Nucleic Acids: Main Characteristics and Clinical Application. Int. J. Mol. Sci., 21.

40. Murchie,T.J., Kuch,M., Duggan,A.T., Ledger,M.L., Roche,K., Klunk,J., Karpinski,E., Hackenberger,D., Sadoway,T., MacPhee,R., et al. (2021) Optimizing extraction and targeted capture of ancient environmental DNA for reconstructing past environments using the PalaeoChip Arctic-1.0 bait-set. Quaternary Research, 99, 305–328.

41. Brandhagen,M.D., Loreille,O. and Irwin,J.A. (2018) Fragmented nuclear DNA is the predominant genetic material in human hair shafts. Genes (Basel), 9.

42. Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. Nat. Rev. Genet., 11, 499–511.

Chapter 2. Methods to make probes for public health applications

Whole genome sequencing of pathogens is important to understand the pathogens strain type, drug-resistance and immune evasion phenotypes. To obtain the DNA, pathogens are isolated and cultured from the samples. For rapid identification and characterization of pathogens, direct sequencing of clinical samples is needed. However, direct sequencing of clinical samples is not feasible due to presence of low fraction of pathogen DNA in vast amount of human and other commensal microbiome DNA isolated from the clinical samples. Therefore, we developed the whole genome enrichment (WGE) approach using the CNER probes to enrich the pathogen genome to directly sequence them from samples.

In the first section of this chapter, I demonstrate the CNERs-WGE approach for sequence tuberculosis causing bacteria, *Mycobacterium tuberculosis*. I discuss how the CNERs-WGE method can be used for genomic epidemiology of difficult to grow pathogens.

In the second section of this chapter, I demonstrate, using *Toxoplasma gondii* as an example, how the CNER-WGE method can be used to enrich large genomes parasites from environmental and food samples. I discuss the application of the CNERs-WGE method for foodborne pathogen screening programs.

Chapter 2.1. A hybridization capture approach for pathogen genomics

Abstract

Genomic epidemiology uses pathogens' whole-genome sequence to understand and manage the spread of infectious diseases. Whole-genome data can be used to monitor outbreak and cluster formation, to identify cross-community transmissions, and to profile drug resistance and immune-evasion. Typically, pathogens are cultured from clinical samples to obtain DNA for sequencing to generate whole-genome data. However, culture-independent diagnostic methods are needed for difficult-to-grow pathogens and for rapid pathogen genomics. Whole-genome enrichment (WGE) using targeted DNA sequencing enables direct sequencing of clinical samples without culturing pathogens. However, the cost of enrichment baits limits the utility of this method for large scale genomic epidemiology. We developed a cost-effective method named Circular Nucleic acid Enrichment Reagent synthesis (CNERs) to generate whole-genome enrichment probes. We demonstrated the method by producing probes *for Mycobacterium tuberculosis* which we used to enrich *M. tuberculosis* DNA that had been spiked at concentrations as low as 0.01% and 100 genome copies against human DNA background to 1225-fold and 4636-fold. Further, we also enriched DNA from different *M. tuberculosis* lineages and M. bovis and demonstrated the utility of the WGE-CNERs data for lineage identification and drug-resistance characterization using an established pipeline. The CNERs method for whole-genome enrichment will be a valuable tool for genomic epidemiology of emerging and difficult-to-grow pathogens.

Introduction

Genomic epidemiology uses pathogens' whole-genome data to understand and manage the spread of infectious diseases (1–3). Genome sequences can be used to predict pathogens' phenotypes including virulence factors, drug-resistance markers, and other factors used in immune-evasion (4). Whole-genome data also is used to identify phylogenetic relationships at a higher resolution than multi-locus markers, allowing fine-scale relationships between pathogenic strains. Genomes of several pathogens have been used to study small outbreaks (5) and to track strain prevalence at the national level (1).

The 2019 SARS-CoV2 pandemic demonstrated the potential of genomic epidemiology in real time disease monitoring. Since 2019, ~15 million SARS-CoV2 clinical genomes have been sequenced (www.gisaid.org) to understand transmission dynamics and track viral evolution. Genomic data has been used to identify the origin, mode of introduction of new infection and cluster formation, and to reveal how clinical isolates are related in space and time (6–8). Further, genomic data characterized viral evolution in response to clinical interventions (9, 10). The SARS-CoV2 pandemic also advanced the technologies and computational capabilities needed to produce and analyze big data for genomic epidemiology (11). Motivated by the success of SARS-CoV2 genomic surveillance, there is a renewed interest in applying genomic epidemiology for emerging pathogens to monitor, predict outbreaks and recommend disease control measures.

Genomic data are generated by whole-genome sequencing (WGS) of pathogens. WGS requires high quality and reasonable quantity of pathogen's DNA, which is often difficult to obtain from clinical samples (12). Pathogens are generally grown in vitro to make DNA for sequencing (13). Culturing pathogens for DNA is routinely done for bacterial pathogens that are simple to grow (14). However, culturing poses a hurdle for some fastidious bacterial, fungal and parasitic pathogens. Difficult-to-culture pathogens are usually slow growers and require special growth conditions like the presence of host cells (14, 15). Further, culturing steps can alter the diversity that was initially present in the clinical samples (16, 17). Culture independent direct sequencing of clinical samples can overcome many of these challenges (18, 19).

Metagenomic methods have been developed to directly sequence clinical samples like sputum, cerebrospinal fluid, and urine to detect and assemble pathogen genomes in an unbiased fashion (20–23). However, the presence of high amounts of human DNA and commensal microbiome DNA can obscure the detection of trace amounts of pathogen DNA (20, 22). In addition, metagenomic methods often fail to achieve the high sequence coverage required for variant detection, which is needed for phenotypic and phylogenetic characterization. Depletion of human DNA and or enrichment of pathogen DNA have been used in culture-independent diagnostic tests (18–23). Various sample processing methods have also been developed to lyse human cells to remove human DNA before lysing pathogens (19). However, these methods still often fail to achieve the depth of sequencing coverage required for genomic epidemiology due to presence of other microbial DNA.

Whole-genome enrichment (WGE) of pathogens' DNA achieves higher genome coverage in direct sequencing of clinical samples (20). Multiplex PCR using sequence specific primers or cDNA amplification using random primers enable WGE for viral genomes due to their smaller genome size (23–25). Hybridization capture using biotinylated RNA or DNA probes have been used to enrich large genomes of bacterial, fungal, and parasitic pathogens from clinical samples (15–18, 26, 27). However, large scale synthesis of capture probes that cover entire genomes is often prohibitively expensive (20, 27). Current probe synthesis methods are inadequate to meet the large-scale requirement for genomic epidemiology.

We developed the CNER method for large-scale DNA bait synthesis to enrich specific genomic regions (Chapter 1). Here, we describe CNER method to make probes against an entire target pathogen genome. We demonstrate the CNER method to produce DNA baits for whole-genome enrichment (WGE) of *Mycobacterium tuberculosis* (*M. tuberculosis*). We enrich the *M. tuberculosis* DNA from an initial representation of 0.01% spiked in with human DNA to a final representation of >85% using the *M. tuberculosis* WGE-CNERs. We also capture a panel of various *M. tuberculosis* lineages and M. bovis (another species within *M. tuberculosis* complex - MTBC) and several non-tuberculous Mycobacteria (NTMs), demonstrating sensitivity and specificity of WGE-CNERs. Further, we also show the utility of the WGE data generated using CNERs for lineage identification and drug-resistance characterization.

Materials and Methods

DNA samples and library preparation

The California Department of Public Health's Microbial Diseases Laboratory kindly provided mycobacterial genomic DNA (gDNA). We prepared NGS libraries using NEB Ultra II FS kit by following kit's manual.

For the proof-of-concept experiment, we intentionally spiked *M. tuberculosis* H37Rv libraries with unique dual indices at 0.01%, 0.1%, 1% and 10% expected representation with human libraries (prepared with NA12878 gDNA). We sequenced the contrived mixture of *M. tuberculosis* and human libraries to confirm the proportion of *M. tuberculosis* libraries before capture.

To determine sensitivity, we spiked in 10 fg – 95.2 ng (8 times 1:10 serial dilution) of *M. tuberculosis* H37Rv gDNA corresponding to 2×10^0 – 2×10^7 *M. tuberculosis* genome copies with 54 ng of human gDNA (NA12878) and prepared NGS libraries using NEB Ultra II FS kit with following modifications. After fragmentation and adapter ligation, we split the adapter ligated DNA into two aliquots (corresponding to 1×10^0 – 1×10^7 *M. tuberculosis* genome copies per library) and amplified them for 8 cycles with NEB Q5 master mix with two sets of unique dual indices. We sequenced all 16 libraries before capture experiments to determine the *M. tuberculosis* proportion in each library.

For the specificity test, we spiked in 140 fg – 142.8 fg (4 times 1:10 serial dilution corresponding to 3×10^1 – 3×10^4 *M. tuberculosis* genome copies) of four *M. tuberculosis* lineages (*M.tb* Indo-Oceanic – Lineage 1, East-Asian – Lineage 2, East-African-Indian – Lineage 3 and Euro-American – Lineage 4), *M. bovis* and three NTMs (*M. abscessus, M. fortuitum* and *M. porcinum*) with 54 ng of human gDNA (NA12878) and prepared NGS libraries using NEB Ultra II FS kit with following modifications. After fragmentation and adapter ligation, we split the adapter ligated DNA into three aliquots (corresponding to 1×10^1 – 1×10^4 *M. tuberculosis* genome copies per library) and amplified them for 12 cycles with NEB Q5 master mix with three sets of unique dual indices. We sequenced all 96 libraries before capture experiments to determine the mycobacteria proportion in each library.

M.tb WGE-CNERs generation

We generated WGE-CNERs to enrich mycobacteria as described in Fig. S1. We sheared ~286 ng of *M. tuberculosis* H37Rv reference gDNA using Covaris in microTUBE15 for 250s with peak power at 50, 30% duty factor and 50 cycle bursts at 23°C. We denatured 100 ng of sheared gDNA at 95°C for 3 min and snap cooled on ice block. Separately, 100 pmol of bridge oligo (5'–GCGCGATCAAGCTTTTTTTTTTTTTTTTTTTTT–3') annealed with splint oligos (5'-NNNNNNNNAAAAAAAAAAA–3' and 5'–GCTTGATCGCGCNNNNNNNN–3') by denaturing at 95°C for 3 min and cooling to 12°C with 0.1°C/s ramp speed. Sheared denatured gDNA mixed with 35 pmol of annealed bridge/splint oligos and ligated in 1X T4 DNA ligase buffer at 37°C for 1 h followed by 25°C for 3 h and denatured at 95°C for 3 min. We amplified the circularized genomic fragments as described in Chapter 1 and digested the RCA products using 50U of HindIII enzyme.

Enrichments and sequencing

For proof-of-concept experiment we hybridized 100 – 300 ng of four contrived mixtures of *M. tuberculosis* and human libraries with 50 ng of *M. tuberculosis* WGE-CNERs at 65°C for 19.5 h. For sensitivity test, we either hybridized 100 ng of 8 individual libraries of various *M. tuberculosis* copy numbers with 25 ng of WGE-CNERs or pooled 25 ng each of the 8 libraries and hybridized the pool with 50 ng of WGE-CNERs.

For specificity test, we pooled two sets of 11 ng each of the *M. tuberculosis* H37Rv, four *M. tuberculosis* lineages, *M. bovis* and three NTM libraries from the same copy number mixtures of $1 \times 10^1 – 1 \times 10^4$ copies. For the hybridization temperature experiment, we captured the pools with 25ng of WGE-CNERs at 55°C and 60°C for 19.5 h. For the hybridization time experiment, we captured the pools with 25ng of

WGE-CNERs at 65°C for 1 h and 4 h. For captures at 65°C, we pooled 12.5 ng each of the 4 lineages, *M. bovis* and 3 NTMs from same copy number libraries.

For all capture experiments, we enriched the captured library on streptavidin beads as described in Chapter 1. We amplified post-capture libraries with 2X Kapa HiFi PCR mix for 17 cycles and purified the libraries using 1.2x SPRI beads. Post-capture libraries were pooled in equimolar ratio and sequenced in the Illumina NextSeq with a PE 2x75 kit. For all experiments we sequenced ~50k–100k raw read pairs. We sequenced ~3M raw read pairs for each of the libraries captured at 65°C for 19.5 h.

Data analysis

We used cutadapt to remove adapter sequences and mapped the trimmed reads to *M. tuberculosis* H37Rv reference (NC_000962.3) using bwa mem. We used samtools rmdup to remove duplicate reads. We used samtools and bedtools to determine the percent mapped reads and genome coverage. We used custom python scripts to plot the metrices. We performed a nonparametric Mann-Whitney U rank test for comparing coverage metrices between different hybridization temperatures and times.

For variant detections, we used HaplotypeCaller from the GATK package with -ERC GVCF and -ploidy 1 options to individually call variants on each sample from 1,000- and 10,000-copy numbers captured at 65°C for 19.5 h. We used CombineGVCFs and GenotypeGVCFs with default options to combine VCFs from five *M. tuberculosis* lineages and *M. bovis* of same copy number. We used vcftools with --max-alleles 2 --remove-indels options to remove indels and filter for biallelic variants. We also used --min-meanDP 5 to filters for read depth. We used bcftools stats to determine the genotype concordance between variants from 1,000- and 10,000-copy number samples.

Results

We previously demonstrated the CNER method to generate DNA baits against specific genomic target regions (Chapter 1) which can be adopted to make probes against entire genome. To generate WGE-CNER, gDNA of either a target pathogen or related taxa is fragmented (Fig. S1) and circularized by splint ligation using a bridge adapter. The bridge adapter contains an upper oligo with a rare cutter restriction enzyme recognition site (RES) and oligo-dT sequences; the bottom oligo is complimentary to the upper oligo with degenerate nucleotides at both ends (Fig. S1). These degenerate nucleotides randomly compliment the ends of target gDNA to facilitate splint ligation of the upper oligo. Ligation of the upper oligo both circularizes and incorporates RES and oligo-dT sequences in the target gDNA regardless of their sequences (Fig. S1). Circularized templates are then amplified by rolling circle amplification (RCA) and digested as described in Chapter 1 to generate double stranded CNERs. WGE-CNERs can be used as baits to capture whole genomes of target species and related taxa.

For this study, we sheared ~ 286 ng of *M. tuberculosis* H37Rv gDNA (Fig. S2A) that generated 60% of the population with 100 bp mean size and 37.5% with 202 bp mean size (Fig. S2B). We circularized 100 ng sheared gDNA with a bridge adapter. RCA amplification of circularized templates yielded 4,760 ng of ~37kB mean size high molecular weight DNA (Fig. S2C). HindIII restriction digestion of RCA products generated 4,640 ng of monomeric CNERs with an average size of 131 bp (Fig. S2D). In a separate experiment, we tested 50 ng sheared gDNA as input template that generated 1.8 µg CNERs. Thus, we estimate that ~50 ng of sheared gDNA can produce ~2 µg WGE-CNERs.

CNERs efficiently enrich whole genomes of *M. tuberculosis*

We made four contrived mixtures of *M. tuberculosis* and human NGS libraries with unique Illumina dual indices, sequenced before-enrichment and confirmed the expected *M. tuberculosis* representations (Table 2.1.1). Sequencing after-enrichment using *M. tuberculosis* WGE-CNERs yielded 84 – 99% *M. tuberculosis* data, representing a 9.9 – 1225.3-fold enrichment which varied based on initial proportions (Table 2.1.1). Correspondingly the human libraries were 0.01 – 0.16-fold depleted to a final representation of 1 – 15.5%.

To test the lowest genome copies that *M. tuberculosis* WGE-CNERs can enrich, we spiked *M. tuberculosis* H37Rv gDNA equivalent to 1 × 10^0 – 1 × 10^7 genome copies with 9 million copies of human genome. Sequencing before-enrichment produced 0 – 52% of reads uniquely mapping to the *M. tuberculosis* H37Rv reference genome (NC_000962.3, Table S1) and is consistent between two PCR replicates (cyan squares and circles in Fig. 2.1.1A). Percent mapped reads exponentially increased corresponding to the 10-fold increase in the genome copies. We enriched the mixtures either individually for each copy-numbers or pooled all mixtures before enrichment. Percent unique mapped reads exponentially increased after-enrichment from 0.1% to 17.7% for 1 × 10^0 – 1 × 10^3 copies and plateaued at ~45% for 1 × 10^4 – 1 × 10^6 copies and reached ~66% for 1 × 10^7 copy mixture (pink squares and dots in Fig. 2.1.1A, Table S1). Percent unique mapped reads differed 2.5-fold between individual and pooled captures of the same copy mixtures for 1 × 10^0 – 1 × 10^3 copies and about 6% for 1 × 10^4 – 1 × 10^7 copy mixtures (Table S1).

Table 2.1.1. WGE of M. tuberculosis from contrived mixture with human libraries.

Mixture	*M. tb* before Capture (expected)	*M. tb* before Capture (actual)	*M. tb* after Capture	*M. tb* fold-enrich-ment	Human before Capture (expected)	Human before Capture (actual)	Human after Capture	Human fold-depletion
1	10.000%	9.983%	98.974%	9.91	90.000%	90.017%	1.026%	0.01
2	1.000%	1.249%	98.731%	79.07	99.000%	98.751%	1.269%	0.01
3	0.100%	0.153%	91.415%	599.32	99.900%	99.847%	8.585%	0.09
4	0.010%	0.069%	84.450%	1,225.28	99.990%	99.931%	15.550%	0.16

We determined the fold-enrichment as the ratio between before- and after-enrichment percent unique mapped reads (Fig. 2.1.1B). Fold enrichment for one-copy mixture is unreliable due to inconsistency in percent mapped reads. Enrichment of ten-copy mixture produced 4,717x fold-enrichment for individual and 12,272x for pooled capture experiments. One hundred and 1,000-copy mixtures produced 117x - 375x (average 241x) fold enrichments which differed ~1.5-fold between individual and pooled captures. The fold-enrichment exponentially decreased from 200x to 1.3x with increasing copies for 1 × 10^4 – 1 × 10^7 mixtures that differed ~10% between individual and pooled capture experiments of the same copy numbers (Table S1).

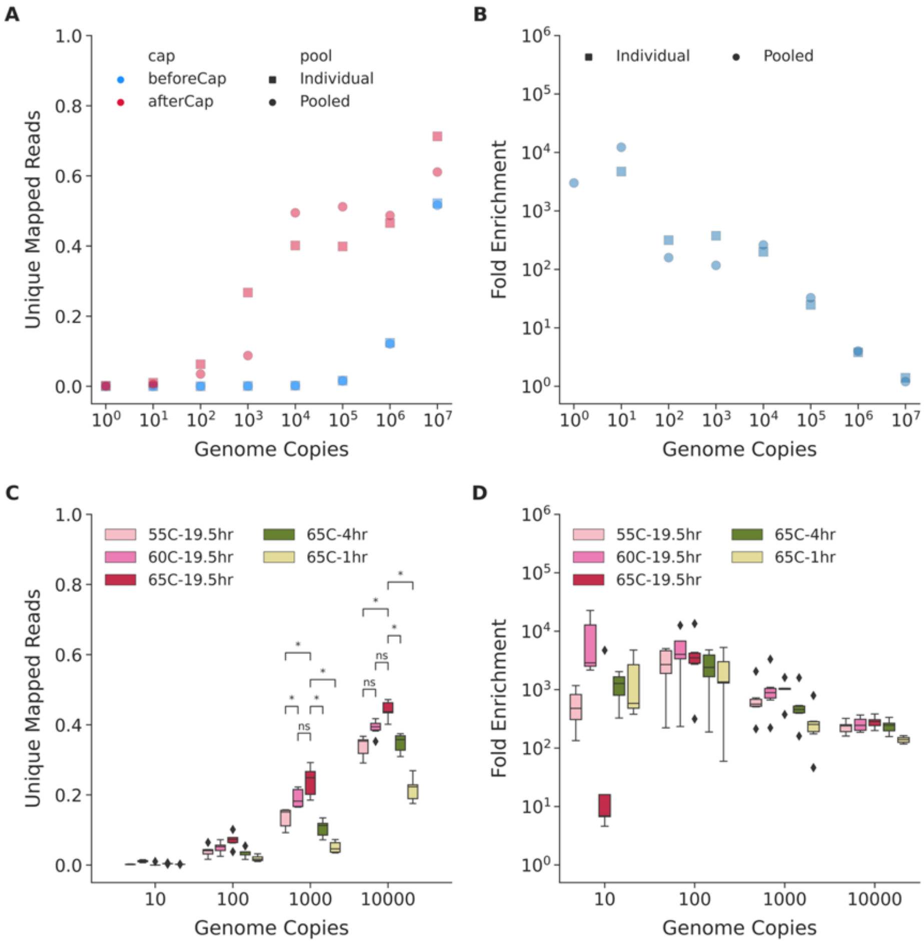


Figure 2.1.1 CNERs efficiently enrich MTBC DNA spiked in with human DNA.

(A) *Percentage of unique mapped reads before (cyan) and after (pink) enrichment with CNERs for individually (squares) and pooled (dots) libraries with 1 × 10^1 – 1 × 10^7 M. tuberculosis genome copies.* ***(B)*** *Fold enrichment which is the ratio between after- and before-enrichment unique mapped reads for individual and pooled captures experiments shown in panel A.* ***(C)*** *Box plots of percentage of unique mapped reads after-enrichment at indicated hybridization temperatures and times for the five M. tuberculosis lineages and M. bovis.* ***(D)*** *Fold enrichment at indicated hybridization temperatures and times. Asterisks denote statistical significance tested by Mann-Whitney Wilcoxon test.*

Pairwise comparison of normalized coverage of 100 bp bins across the genome show that the coverage after CNERs enrichment is highly correlated for samples with 10,000 or more copies (Pearson's r = 0.90 - 0.93, Fig. S3). The normalized coverage shows that certain regions in the genome are preferentially enriched at given sequencing depth when 1000 or less copies are present (Fig. S3), eventually the pairwise correlation for these samples decreases against higher copy samples. The average Pearson's r is 0.87, 0.70, 0.37 and 0.095 for 1000, 100, 10 and 1 copy samples against higher copy samples.

CNERs specifically enrich for TB lineages and MTBC species

We asked whether the CNERs made using the *M. tuberculosis* - H37Rv gDNA can enrich different species and lineages of MTBC and NTM. We made libraries for 1 × 10^1 – 1 × 10^4 genome copies of four *M. tuberculosis* lineages (Indo-Oceanic – Lineage 1, East-Asian – Lineage 2, East-African-Indian – Lineage 3 and Euro-American – Lineage 4), *M. bovis* and three NTM species (*M. abscessus, M. fortuitum* and *M. porcinum*) mixed with human DNA. We generated ~153k read pairs for each library before enrichment which resulted in 0 – 0.2% unique mapped reads which varied based on the genome copies as expected (Table S1).

We pooled same copy number libraries of different taxa and captured them at 55°C, 60°C and 65°C to test the effect of hybridization temperature on capture efficiency. We sequenced ~53k read pairs for each libraries after-enrichment which generated 0 – 47.5% unique mapped reads when mapped to the NC_000962 reference (shades of pink in Fig. 2.1.1C, Table S1). Similar to the H37Rv captures, unique mapped reads increased with increasing copy number for the *M. tuberculosis* lineages and *M. bovis* (Fig. 2.1.1C) but did not improve the mapped reads for the NTMs (Fig. S4, Table S1). The NTMs produced <7% unique mapped reads (Fig. S4). The low percentage of unique mapped reads for NTM samples might be due to either poor capture of NTM DNA by *M. tuberculosis* WGE-CNERs or poor mapping of NTM reads to the NC_000962 reference. To test this, we mapped after-enrichment reads of the NTMs to three NTM references (Fig. S4) which produced ~9.3% unique mapped reads on average. The mapped reads slightly differed between three NTM references. Mapping to the individual NTM references showed that the low percent mapped reads are due to poor capture of NTM genomes by *M. tuberculosis* CNERs rather than poor mapping to *M. tuberculosis* reference. Five-fold differences between *M. tuberculosis* and NTMs in the after-enrichment unique mapped reads demonstrate that the *M. tuberculosis* WGE-CNERs are specific to the MTBCs and does not enrich NTM genomes. The modest improvements in the percent mapped reads for NTMs between before- and after-enrichments indicate that *M. tuberculosis* WGE-CNERs might enrich small portions of conserved regions in the genomes of all mycobacterial species.

For MTBC, hybridization at 65°C produced on average 44.2% and 23.9% unique mapped reads compared to 39.1% and 19.0% at 60°C and 33.8% and 13.5% at 55°C for the 10,000- and 1,000-copy mixtures respectively. For the 10- and 100-copy mixtures, the percent unique mapped reads were <10% and did not significantly differ between different hybridization temperatures (Fig. 2.1.1C, S5 and Table S1). The fold-enrichment decreased with increasing copy numbers similar to the H37Rv captures (Fig. 2.1.1D, S6 and Table S1). WGE of 10,000-copy mixture resulted on average 250x fold-enrichment.

We are interested in reducing the overnight hybridization to 4 or less hours to make a one-day enrichment protocol for rapid clinical diagnostics. We enriched the MTBC mixture for 1 h or 4 h hybridizations at 65°C. Hybridization for 1 h produced on average 21.6% and 5.0% unique mapped reads compared to 34.8% and 10.5% for 4 h and, 44.2% and 23.9% for 19.5 h at 65°C (data from the temperature experiments) for the 10,000- and 1,000-copy mixtures respectively (green and yellow bars in Fig. 2.1.1C and Table S1). Hybridization duration does not significantly change the percent unique mapped reads for the 10- and 100-copy mixtures.

CNERs enrichment produce high coverage MTBC genomes

We deeply sequenced after-enrichment samples captured at 65°C for 19.5 h to analyze genomic coverage. We blasted 50,000 raw reads using local BLASTn against the NCBI nucleotide (nt) database and analyzed the results using Metagenomic Analyzer (MEGAN) software to independently check the reads originating from MTBC. MEGAN assigned 0.5%, 12.5%, 37.1% and 53.8% of reads to *M. tuberculosis* taxa on average for the five *M. tuberculosis* lineages and *M. bovis* samples for the 10-, 100-, 1000- and 10000-copy mixtures respectively (Fig. S7). The percent of taxa assigned as *M. tuberculosis* is consistent with the percent mapped reads to the NC_000962 reference (Fig 2.1.1C).

We subsampled to three million raw reads corresponding to ~102X of *M. tuberculosis* genomic bases that produced 0%, 0.7%, 7.1% and 28.2% unique mapped reads with 81.2%, 93.7%, 82.1% and 53.9% duplication rate for the four copy mixtures. We measured the coverage at each genomic position and plotted the percentage of the genome covered with x or more unique reads.

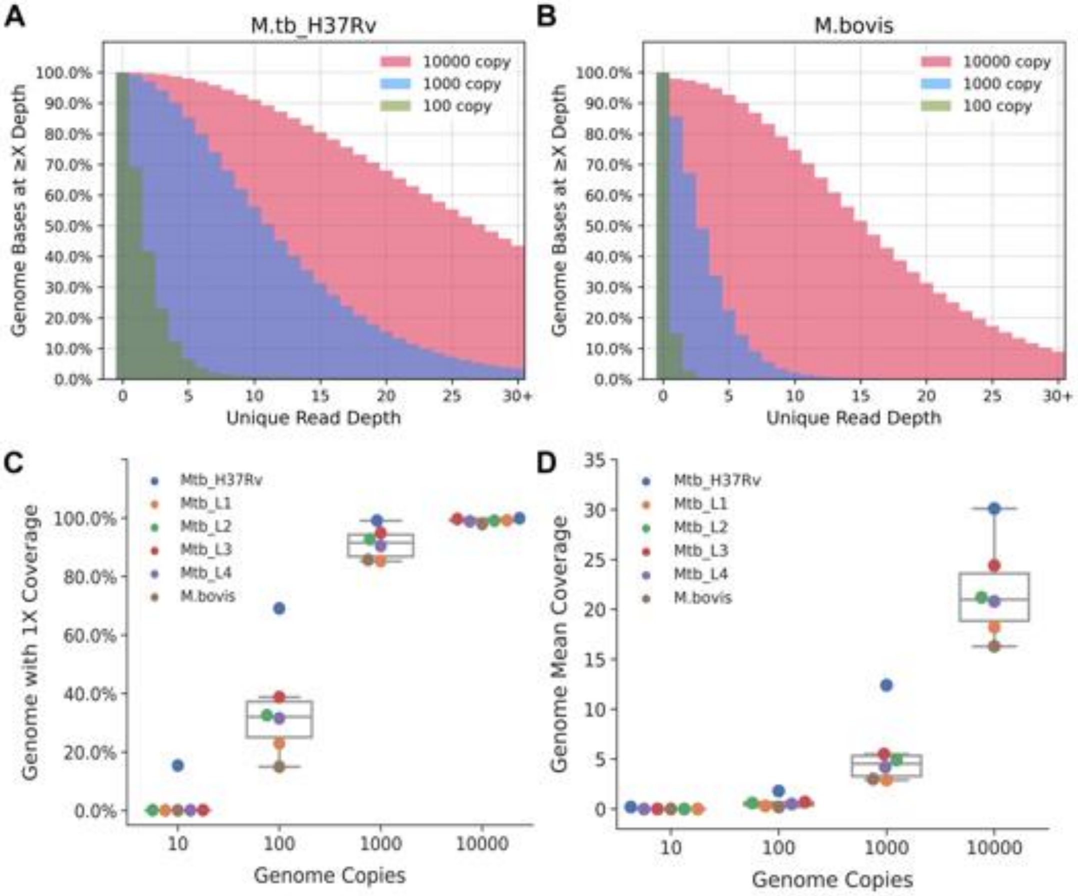


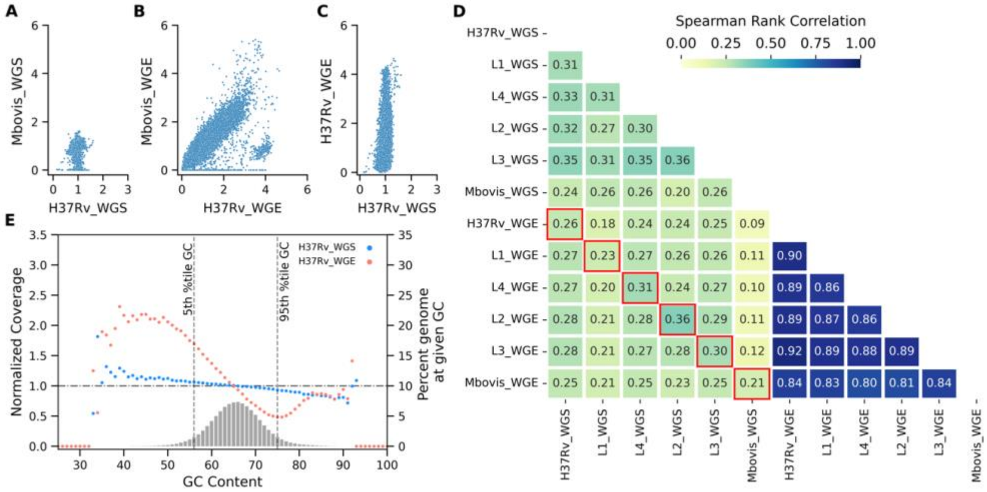
***Figure 2.1.2 Breadth and depth of genome coverage using WGE-CNERs data**. Overlapping histogram of percent of genome with X or more unique read depth from three million reads of WGE-CNERs data for 10,000-copy (pink), 1,000-copy (cyan) and 100-copy (green) mixtures of M. tuberculosis H37Rv **(A)** and M. bovis **(B).** Box plots with overlapping swarm plots of percent of genome with 1X or more read coverage **(C)** and genome unique mean coverage **(D)** resulting from three million reads of WGE-CNERs data five M. tuberculosis lineages and M. bovis (color dots) at indicated copy numbers.*

The coverage for *M. tuberculosis* H37Rv (Fig. 2.1.2A) for a given copy mixture is slightly better compared to *M. bovis* (Fig. 2.1.2B) and other *M. tuberculosis* lineages (Fig. S8). We analyzed the breadth of coverage by looking at the *M. tuberculosis* genome bases covered with at least one read (1X coverage). For the H37Rv, the 1X coverage for the four copy mixtures is 15.4%, 69.1%, 99.1% and 99.9% (blue dots in Fig. 2.1.2C) compared to 0%, 28.1%, 89.8% and 99.0% on average of the other four lineages and *M. bovis* (Fig. 2.1.2C). At the given three million raw read pairs, the unique mean coverage depth for H37Rv is 0.2, 1.8, 12.4 and 30.1 (blue dots in Fig. 2.1.2D) compared to 0, 0.5, 4.1 and 20.2 for the other five MTBC samples (Fig. 2.1.2D).

We generated shotgun WGS data using the same mycobacterial DNA samples that produced 68x average genome coverage from three million raw read pairs. We normalized the coverage for 100bp genomic bins to account for the differences in the absolute coverage to compare WGS with WGE-CNERs. WGS resulted in uniform coverage across the 100bp bins (normalized coverage closer to 1, Fig. 2.1.3A and S9) that differed between lineages and species consistent with expected genomic differences among MTBC.

WGE-CNERs data also reproduced the difference between the lineages and species (Fig. 2.1.3B and S10). However, the normalized coverage for different genomic loci varied up to 5-fold within a sample (Fig. 2.1.3B, 3C and S11) compared to a more uniform coverage in WGS (Fig. 2.1.3C and S9). Pairwise comparisons of WGS coverage between the MTBCs show a weak correlation with an average Spearman rho of 0.29 (upper triangle in Fig. 3D), but strong correlation between the WGE-CNERs experiments with average Spearman rho of 0.86 (lower triangle in Fig.

2.1.3D). The genomic differences between the MTBCs compounded with uneven coverage in WGE-CNERs resulted in lower correlation between WGS and WGE-CNERs for different mycobacteria (center block in Fig. 2.1.3D). However, WGS and WGE-CNERs for the same mycobacteria were correlated with an average Spearman rho of 0.28 (center diagonal, highlighted in Fig. 2.1.3D), similar to the WGS correlations.



	H37Rv_WGS	L1_WGS	L4_WGS	L2_WGS	L3_WGS	Mbovis_WGS	H37Rv_WGE	L1_WGE	L4_WGE	L2_WGE	L3_WGE	Mbovis_WGE
H37Rv_WGS												
L1_WGS	0.31											
L4_WGS	0.33	0.31										
L2_WGS	0.32	0.27	0.30									
L3_WGS	0.35	0.31	0.35	0.36								
Mbovis_WGS	0.24	0.26	0.26	0.20	0.26							
H37Rv_WGE	0.26	0.18	0.24	0.24	0.25	0.09						
L1_WGE	0.27	0.23	0.27	0.26	0.26	0.11	0.90					
L4_WGE	0.27	0.20	0.31	0.24	0.27	0.10	0.89	0.86				
L2_WGE	0.28	0.21	0.28	0.36	0.29	0.11	0.89	0.87	0.86			
L3_WGE	0.28	0.21	0.27	0.28	0.30	0.12	0.92	0.89	0.88	0.89		
Mbovis_WGE	0.25	0.21	0.25	0.23	0.25	0.21	0.84	0.83	0.80	0.81	0.84	

Figure 2.1.3 Correlation of normalized coverage at 100 genomic bins between WGS and WGE-CNERs data.

*Scatter plot of normalized coverage between M. tuberculosis H37Rv and M. bovis generating by WGS **(A)** and WGE-CNERs **(B). (C)** Scatter plot of normalized coverage from WGS vs WGE-CNERs for M. tuberculosis H37Rv 10,000-copy sample. **(D)** Heat map of Spearman rank correlations of pairwise comparisons of normalized coverage between WGS and WGE for five M. tuberculosis lineages and M. bovis. **(E)** Scatter plot of normalized coverage (primary Y-axis) across G+C bins and histogram (secondary Y-axis) of percentage of G+C bins plotted for WGS (cyan) and WGE-*

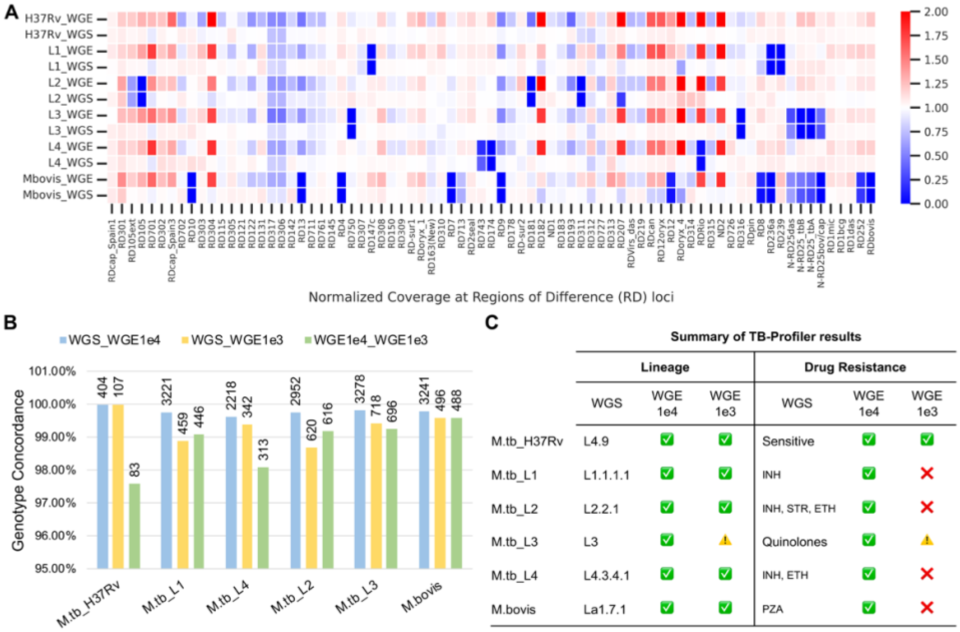
CNERs (orange) data for M. tuberculosis H37Rv. Horizontal line show normalized coverage at 1 and two vertical lines show the 5th and 95th percentile G+C bins of the genome.

We determined the normalized coverage across G+C bins using Picard tool's *CollecGCBias* to check the effect of G+C content on coverage. For 90% of genome with G+C content of 57% – 75%, centered at 65% mean, the coverage differed ~6% from the mean coverage in WGS but differed ~55% from the mean coverage in the WGE-CNERs (Fig. 2.1.3E and S12). For the 10% of the genome with extreme G+C (<57% and >75%), the normalized coverage varied 0.5 – 2.5-fold in the CENRs-WGE (Fig. 2.1.3E and S12).

Detection of lineages and resistance determinants from WGE-CNERs data

Regions of Difference (RD) loci are genome wide small insertion deletions specific to individual MTBC samples that are used for clinical strain-typing (28). We plotted the normalized coverage at RD loci as heatmap to determine the coverage. The WGS data identified RD deletions specific to each MTBC where the coverage is zero (blue boxes in Fig. 2.1.4A). The 10,000-copy WGE-CNERs data also identified the RD loci deletions where the coverage is zero (blue boxes in Fig. 2.1.4A and S13A), but without G+C based coverage normalization, the heatmap showed many RD loci with two-fold coverage (Fig. S13A). We normalized the coverage based on G+C contents to eliminate most of coverage unevenness in WGE-CNERs data. Yet, some RD loci appear to have two-fold coverage only in the WGE-CNERs (red boxes in Fig. 2.1.4A).

The 1,000-copy mixtures WGE-CNERs data also identified RD loci deletions (Fig. S13B). Due to G+C coverage difference, the WGE-CNERs data can only be used to discern deletions which are consistent with the WGS data, but not tandem duplications.



Summary of TB-Profiler results

	Lineage			Drug Resistance		
	WGS	WGE 1e4	WGE 1e3	WGS	WGE 1e4	WGE 1e3
M.tb_H37Rv	L4.9	✓	✓	Sensitive	✓	✓
M.tb_L1	L1.1.1.1	✓	✓	INH	✓	✗
M.tb_L2	L2.2.1	✓	✓	INH, STR, ETH	✓	✗
M.tb_L3	L3	✓	⚠	Quinolones	✓	⚠
M.tb_L4	L4.3.4.1	✓	✓	INH, ETH	✓	✗
M.bovis	La1.7.1	✓	✓	PZA	✓	✗

Figure 2.1.4 WGE-CNERs data can be used to genotype MTBC.

(A) *Heatmap of normalized coverage from WGS and WGE data for the five M. tuberculosis lineages and M. bovis at Regions of Difference (RD) loci sorted by their genomic location.* ***(B)*** *SNP genotype concordance between WGS vs WGE-1e4 (cyan), WGS vs WGE-1e3 (yellow) and WGE-1e4 vs WGE-1e3 (green). Labels on top of the bar show the number of overlapping genotyped sites.* ***(C)*** *Summary of the TB-Profiler results. A green tick mark indicates agreement of TB-Profiler results between WGS and WGE for lineage and drug-resistance pattern. A red x-mark or caution indicates misclassification or no-classification of lineage and drug-resistance pattern in the WGE-1e3 data compared to WGS.*

Sequencing data are used for strain typing and drug resistance profiling in genomic epidemiology. We used GATK HaplotypeCaller with ploidy=1 option to call variants, removed indels and filtered for minimum read depth of five reads in the WGS and WGE data. We genotyped on average 5,444 genome wide positions using the WGS data, 5,416 positions using the 10,000-copy and 4,188 positions using the 1,000-copy mixtures WGE-CNERs data. We calculated the genotype concordance as the percentage of genotypes (both reference and alternative alleles) that matched between the WGS and WGE_CNERs data over the total number of genotyped positions. Among the 2,552 overlapping positions in all five *M. tuberculosis* lineages and *M. bovis* between WGS and 10,000-copy WGE-CNERs data, 99.80% concur (Fig. 4B, cyan bars), 99.35% of 457 overlapping sites concur between WGS and 1,000-copy WGE-CNERs data, (Fig. 2.1.4B, yellow bars) and 98.81% of 440 overlapping sites concur between 1,000-copy and 10,000-copy WGE-CNERs data (Fig. 2.1.4B, green bars).

We find on average 36 variants (14 - 45) per sample identified among the 34 drug-resistance conferring genes that concur 100% between the WGS and 10,000-copy mixtures WGE-CNERs data. To characterize the variants in the drug-resistance conferring genes and to identify the drug-resistance pattern, we used TB-Profiler (29) web tool. The WGS and 10,000-copy WGE-CNERs data correctly identified the expected lineages, both of which are also matched for all six samples using TB-Profiler (Fig. 2.1.4C and Table S2). Lineages for five out of six samples are also correctly identified using the 1,000-copy WGE-CNERs data. Lineage is not determined for the L3 sample using the 1,000-copy WGE-CNERs data, which might be due to low-coverage. TB-Profiler also identified the drug-resistance pattern using variants identified in the WGS and 10,000-copy WGE-CNERs. Both data identified that the *M. tuberculosis* H37Rv is sensitive to all drugs; L1 is resistant to isoniazid (INH) due to Ser315Thr missense mutation in KatG; L2 is resistant to three first line drugs (isoniazid-INH, streptomycin-STR, and ethionamide-ETH); L3 is resistant to all Quinolones due to Ser91Pro missense mutation in Gyrase A; L4 is resistant to isoniazid and ethionamide; and *M. bovis* is resistant to pyrazinamide (PZA) due to His57Asp missense mutation in PncA (Fig. 2.1.4C and Table S2). Due to the coverage cutoff of 10 reads in the TB-Profiler pipeline to assign variants, four samples are misclassified as sensitive to all drugs and L3 sample is misclassified as resistant to ethionamide (ETH) using the 1,000-copy WGE-CNERs data (Fig. 2.1.4C and Table S2).

Discussion

We demonstrated that the CNERs method can generate microgram quantities of WGE baits, which were used to enrich MTBC DNA for genomic analyses. We spiked a wide range of MTBC genome copies to a constant amount of human DNA to mimic clinical sputum samples. Our results illustrated that the WGE-CNERs can enrich *M. tuberculosis* DNA as low as 0.01% in the initial sample and 100 - 10,000 absolute copies of MTBC genomes from a vast majority of human DNA background. We showed that the breadth and depth of genome coverage using WGE-CNERs depended on the copy number in the initial sample and after-enrichment sequencing depth as previously observed for direct sequencing of clinical samples (16–18, 30–33). Further, we also showed that short-duration (1 - 4 h) hybridization using WGE-

CNERs can detect 1,000 or more bacilli, and overnight (16 – 20 h) hybridization can detect as low as 100 tuberculosis bacilli. *M. tuberculosis* detection threshold for the sputum acid fast smear test is 5,000 - 10,000 bacilli/ml (34) and for the Xpert MTB method is ~100 bacilli /ml (34, 35). The detection threshold may vary based on the initial volume of sputum sample processed in these methods which must be noted when comparing detection threshold of WGE-CNERs enrichment with these methods. Previous studies find higher concordance in drug-resistance genotype between direct sequencing and WGS after culturing the isolates (18, 31, 32). However, studies also identified higher genetic diversity and hetero-drug resistance from direct sequencing of clinical samples which are lost after culturing (16, 30). We demonstrated the WGE-CNERs method using contrived mixtures of mycobacterial DNA isolated from pure culture mixed with human DNA. Though the mixtures mimic clinical specimens with human and pathogen DNA, they were lacking other microbiome DNA present in actual clinical samples. Future work is needed to evaluate how the microbiome DNA present in various clinical samples may affect the WGE-CNERs enrichment efficiency. In addition, further studies are needed to evaluate concordance between predicted drug-resistance patterns between direct sequencing of clinical samples using WGE-CNERs enrichment and sequencing from pure culture isolates. We expect the WGE-CNERs method may have utility in molecular drug susceptibility testing and would reduce time-to-results that are currently a challenge using time-intensive culture-based methods.

We also demonstrated that *M. tuberculosis* CNERs specifically enrich MTBC genomes and poorly enrich NTM genomes similar to the RNA baits previously used (32). NTMs cause pulmonary disease and may coinfect with MTBC in TB endemic regions. Differentiation between MTBC and NTMs is necessary for clinical decisions as both manifest as acid-fast bacilli in clinical smear testing (36, 37). It would be interesting to generate NTM-specific WGE-CNERs for use in combination with *M. tuberculosis* CNERs to expand the existing SNP panel (38) for a unified WGE panel to capture a variety of primary pathogenic Mycobacteria.

WGE is more cost-effective than the shotgun WGS approach to sequence pathogen genomes directly from clinical samples. However, expensive custom-made baits prohibit this approach for large scale sequencing needed for genomic epidemiology. Previous methods of WGE also require longer turnaround time (TAT) due to overnight hybridization, which is critical during epidemics (21). Hybridization time can be shortened for CNERs probes to reduce the TAT. Further, the CNERs method produces microgram quantities of probes that would make it cost-effective than currently available commercial probes. These advantages make the CNERs an alternative for custom made RNA baits for WGE sequencing for genomic epidemiology applications.

The CNERs method has several advantages compared to previously used WGE approaches. Previous approaches used to detect several difficult-to-grow pathogens requires expensive custom-made baits and availability of high-quality reference genome (18, 26, 27). Also, additional probes may be required to capture individual lineages within a species (31). These requirements limit the ability to sequence novel pathogens and new variants. CNERs method does not require prior genome sequence information but requires purified DNA from the pathogen of interest or a closely related species. CNERs can hybridize and enrich homologous sequences with some sequence diversity. Therefore, CNERs can be used to capture lineages and sub-species without the need for customized baits.

Genome size, GC content and large indel variations between strains might be limiting factors when adopting the CNERs method for other pathogens. We observed a five-fold coverage difference between different GC regions in *M. tuberculosis*. Repeat elements and conserved genomic regions among pathogens of same family or phylum might also be other limiting factors when adopting CNERs for pathogens with large genomes. However, we have demonstrated the utility of the WGE-CNERs approach for large parasites by making CNERs against *Toxoplasma gondii* (Chapter 2.2).

The CNERs method described here can be extended to culture-independent diagnostic tests (CIDTs) for other difficult-to-grow pathogens. Current (q)PCR based CIDTs methods offer rapid detection of pathogens but provide only limited detection of antimicrobial resistance (AMR) determinants. CIDTs also fail to provide genotype data for clinical isolates which may impede outbreak surveillance. Furthermore, the laboratories that implement CIDTs for foodborne pathogens tend to skip bacterial isolation and culturing which are required for surveillance networks that utilize WGS pipelines for strain typing (39, 40). We propose WGE-CNERs as an alternative method which can provide the benefits of both culture-based WGS and CIDTs methods. We envision the WGE-CNERs approach adopted not only for rapid detection as a CIDT, but also for routine genomic surveillance to characterize AMR patterns, to detect emerging clinical strains and lineages and to predict outbreaks of a wide range of microbial pathogens.

Data Availability

All raw sequencing data generated for this project are submitted to the NCBI-SRA database under BioProject accession number PRJNA946035. Bioinformatic data processing pipeline and custom python scripts used for making figures are available in the GitHub page:

https://github.com/bsun210/WGE_CNERs_Mtb_pathogen_genomics

Supplemental Material

Supplemental material is available online only.

References

1. Tang P, Croxen MA, Hasan MR, Hsiao WWL, Hoang LM. 2017. Infection control in the new age of genomic epidemiology. Am J Infect Control 45:170–179.

2. Tang P, Gardy JL. 2014. Stopping outbreaks with real-time genomic epidemiology. Genome Med 6:104.

3. Robinson ER, Walker TM, Pallen MJ. 2013. Genomics and outbreak investigation: from sequence to consequence. Genome Med 5:36.

4. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, Posey JE, Gwinn M. 2019. Pathogen genomics in public health. N Engl J Med 381:2569–2580.

5. Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. 2021. Progress and challenges in virus genomic epidemiology. Trends Parasitol 37:1038–1049.

6. Alteri C, Cento V, Piralla A, Costabile V, Tallarita M, Colagrossi L, Renica S, Giardina F, Novazzi F, Gaiarsa S, Matarazzo E, Antonello M, Vismara C, Fumagalli R, Epis OM, Puoti M, Perno CF, Baldanti F. 2021. Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. Nat Commun 12:434.

7. Popa A, Genger J-W, Nicholson MD, Penz T, Schmid D, Aberle SW, Agerer B, Lercher A, Endler L, Colaço H, et al. 2020. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med 12.

8. Dhar MS, Marwal R, Vs R, Ponnusamy K, Jolly B, Bhoyar RC, Sardana V, Naushin S, Rophina M, Mellan TA, Mishra S, et al. 2021. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. Science 374:995–999.

9. Telenti A, Hodcroft EB, Robertson DL. 2022. The Evolution and Biology of SARS-CoV-2 Variants. Cold Spring Harb Perspect Med 12.

10. Safari I, Elahi E. 2022. Evolution of the SARS-CoV-2 genome and emergence of variants of concern. Arch Virol 167:293–305.

11. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast Sample placement on Existing

tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet 53:809–816.

12. Retchless AC, Fox LM, Maiden MCJ, Smith V, Harrison LH, Glennie L, Harrison OB, Wang X. 2019. Toward a global genomic epidemiology of meningococcal disease. J Infect Dis 220:S266–S273.

13. Köser CU, Ellington MJ, Cartwright EJP, Gillespie SH, Brown NM, Farrington M, Holden MTG, Dougan G, Bentley SD, Parkhill J, Peacock SJ. 2012. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 8:e1002824.

14. Tagini F, Greub G. 2017. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. Eur J Clin Microbiol Infect Dis 36:2007–2020.

15. Dennis TPW, Mable BK, Brunelle B, Devault A, Carter RW, Ling CL, Mmbaga BT, Halliday JEB, Oravcova K, Forde TL. 2022. Target-enrichment sequencing yields valuable genomic data for challenging-to-culture bacteria of public health importance. Microb Genom 8.

16. Nimmo C, Shaw LP, Doyle R, Williams R, Brien K, Burgess C, Breuer J, Balloux F, Pym AS. 2019. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture. BMC Genomics 20:389.

17. Lozano N, Lanza VF, Suárez-González J, Herranz M, Sola-Campoy PJ, Rodríguez-Grande C, Buenestado-Serrano S, Ruiz-Serrano MJ, Tudó G,

Alcaide F, Muñoz P, García de Viedma D, Pérez-Lago L. 2021. Detection of Minority Variants and Mixed Infections in Mycobacterium tuberculosis by Direct Whole-Genome Sequencing on Noncultured Specimens Using a Specific-DNA Capture Strategy. mSphere 6:e0074421.

18. Brown AC, Bryant JM, Einer-Jensen K, Holdstock J, Houniet DT, Chan JZM, Depledge DP, Nikolayevskyy V, Broda A, Stone MJ, Christiansen MT, Williams R, McAndrew MB, Tutill H, Brown J, Melzer M, Rosmarin C, McHugh TD, Shorten RJ, Drobniewski F, Speight G, Breuer J. 2015. Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. J Clin Microbiol 53:2230–2237.

19. Doughty EL, Sergeant MJ, Adetifa I, Antonio M, Pallen MJ. 2014. Culture-independent detection and characterisation of Mycobacterium tuberculosis and M. africanum in sputum samples using shotgun metagenomics on a benchtop sequencer. PeerJ 2:e585.

20. Bachmann NL, Rockett RJ, Timms VJ, Sintchenko V. 2018. Advances in clinical sample preparation for identification and characterization of bacterial pathogens using metagenomics. Front Public Health 6:363.

21. Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9–20.

22. Gu W, Miller S, Chiu CY. 2019. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. Annu Rev Pathol 14:319–338.

23. Chiu CY, Miller SA. 2019. Clinical metagenomics. Nat Rev Genet 20:341–355.

24. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, Stremlau M, Berlin A, Gire SK, England E, Moses LM, Mikkelsen TS, Odia I, Ehiane PE, Folarin O, Goba A, Kahn S, Grant DS, Honko A, Hensley L, Happi C, Garry RF, Malboeuf CM, Birren BW, Gnirke A, Levin JZ, Sabeti PC. 2014. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol 15:519.

25. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, Wiley MR, White S, Thézé J, Magnani DM, et al. 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature 546:401–405.

26. Christiansen MT, Brown AC, Kundu S, Tutill HJ, Williams R, Brown JR, Holdstock J, Holland MJ, Stevenson S, Dave J, Tong CYW, Einer-Jensen K, Depledge DP, Breuer J. 2014. Whole-genome enrichment and sequencing of Chlamydia trachomatis directly from clinical samples. BMC Infect Dis 14:591.

27. Clark SA, Doyle R, Lucidarme J, Borrow R, Breuer J. 2018. Targeted DNA enrichment and whole genome sequencing of Neisseria meningitidis directly from clinical specimens. Int J Med Microbiol 308:256–262.

28. Bespiatykh D, Bespyatykh J, Mokrousov I, Shitikov E. 2021. A Comprehensive Map of Mycobacterium tuberculosis Complex Regions of Difference. mSphere 6:e0053521.

29. Phelan JE, O’Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, O’Grady J, McNerney R, Hibberd ML, Viveiros M, Huggett JF, Clark TG. 2019. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med 11:41.

30. Doyle RM, Burgess C, Williams R, Gorton R, Booth H, Brown J, Bryant JM, Chan J, Creer D, Holdstock J, Kunst H, Lozewicz S, Platt G, Romero EY, Speight G, Tiberi S, Abubakar I, Lipman M, McHugh TD, Breuer J. 2018. Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant Mycobacterium tuberculosis Faster than MGIT Culture Sequencing. J Clin Microbiol 56.

31. Goig GA, Cancino-Muñoz I, Torres-Puente M, Villamayor LM, Navarro D, Borrás R, Comas I. 2020. Whole-genome sequencing of Mycobacterium tuberculosis directly from clinical samples for high-resolution genomic epidemiology and drug resistance surveillance: an observational study. The Lancet Microbe 1:e175–e183.

32. Soundararajan L, Kambli P, Priyadarshini S, Let B, Murugan S, Iravatham C, Tornheim JA, Rodrigues C, Gupta R, Ramprasad VL. 2020. Whole genome enrichment approach for rapid detection of Mycobacterium tuberculosis and drug resistance-associated mutations from direct sputum sequencing. Tuberculosis (Edinb) 121:101915.

33. Barbosa-Amezcua M, Cuevas-Córdoba B, Fresno C, Haase-Hernández JI, Carrillo-Sánchez K, Mata-Rocha M, Muñoz-Torrico M, Bäcker C, González-

Covarrubias V, Alaez-Verson C, Soberón X. 2022. Rapid Identification of Drug Resistance and Phylogeny in M. tuberculosis, Directly from Sputum Samples. Microbiol Spectr 10:e0125222.

34. van Zyl-Smit RN, Binder A, Meldau R, Mishra H, Semple PL, Theron G, Peter J, Whitelaw A, Sharma SK, Warren R, Bateman ED, Dheda K. 2011. Comparison of quantitative techniques including Xpert MTB/RIF to evaluate mycobacterial burden. PLoS ONE 6:e28815.

35. Chakravorty S, Simmons AM, Rowneki M, Parmar H, Cao Y, Ryan J, Banada PP, Deshpande S, Shenai S, Gall A, Glass J, Krieswirth B, Schumacher SG, Nabeta P, Tukvadze N, Rodrigues C, Skrahina A, Tagliani E, Cirillo DM, Davidow A, Denkinger CM, Persing D, Kwiatkowski R, Jones M, Alland D. 2017. The New Xpert MTB/RIF Ultra: Improving Detection of Mycobacterium tuberculosis and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing. MBio 8.

36. Maiga M, Siddiqui S, Diallo S, Diarra B, Traoré B, Shea YR, Zelazny AM, Dembele BPP, Goita D, Kassambara H, Hammond AS, Polis MA, Tounkara A. 2012. Failure to recognize nontuberculous mycobacteria leads to misdiagnosis of chronic pulmonary tuberculosis. PLoS ONE 7:e36902.

37. Yoon J-K, Kim TS, Kim J-I, Yim J-J. 2020. Whole genome sequencing of Nontuberculous Mycobacterium (NTM) isolates from sputum specimens of co-habiting patients with NTM pulmonary disease and NTM isolates from their environment. BMC Genomics 21:322.

38. He Y, Gong Z, Zhao X, Zhang D, Zhang Z. 2020. Comprehensive Determination of Mycobacterium tuberculosis and Nontuberculous Mycobacteria From Targeted Capture Sequencing. Front Cell Infect Microbiol 10:449.

39. Tack DM, Ray L, Griffin PM, Cieslak PR, Dunn J, Rissman T, Jervis R, Lathrop S, Muse A, Duwell M, Smith K, Tobin-D'Angelo M, Vugia DJ, Zablotsky Kufel J, Wolpert BJ, Tauxe R, Payne DC. 2020. Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2016-2019. MMWR Morb Mortal Wkly Rep 69:509–514.

40. Ray LC, Griffin PM, Wymore K, Wilson E, Hurd S, LaClair B, Wozny S, Eikmeier D, Nicholson C, Burzlaff K, Hatch J, Fankhauser M, Kubota K, Huang JY, Geissler A, Payne DC, Tack DM. 2022. Changing Diagnostic Testing Practices for Foodborne Pathogens, Foodborne Diseases Active Surveillance Network, 2012-2019. Open Forum Infect Dis 9:ofac344.

Chapter 2.2. Targeted sequencing to detect foodborne pathogens

Abstract

Pathogenic microbes directly or by producing toxins can cause food and waterborne diseases. To control food and waterborne pathogens, surveillance methods screen for pathogens at various stages of food production to consumption. Nucleic acid amplification and biosensor detection methods offer rapid identification of pathogens, but they don’t characterize the pathogens’ characteristics like strain type, virulence traits, or antimicrobial resistance. Whole genome sequencing (WGS) can provide this information. However, pathogens must be isolated and grown for DNA to generate WGS, which is time consuming. We developed the Circular Nucleic acid Enrichment Reagent (CNER) method to make whole genome enrichment (WGE) baits for pathogens. WGE using CNERs will facilitate direct sequencing of pathogens from samples without the need to isolate and grow them. Here, we made WGE-CNERs for *Toxoplasma gondii* to demonstrate the use of the CNER method to make baits for the large protist parasite genomes, using which we detect the parasite by sequencing. We discuss the use of WGE-CNERs to monitor drug-resistant microbes for One Health studies.

Introduction

Food and waterborne diseases are a serious threat to global public health. Foodborne disease can be caused by microbial pathogens including viral, bacterial, fungal and parasitic protist pathogens that can contaminate any stage of food production to consumption (Kirk et al., 2015). Therefore, it is important to screen the food products for microbial pathogens at various stages. Culture based whole genome sequencing (WGS) (Brown et al., 2019) methods and culture independent methods like nucleic acid amplification tests (NAATs), biosensors and antibody based immunological methods, have been used to screen food and waterborne pathogens (Law et al., 2014; Ray et al., 2022). The NAAT and biosensor methods offer rapid identification of pathogens but lack strain-typing and functional characterization like identification of antimicrobial resistance (AMR) (Law et al., 2014; Ray et al., 2022).

Culture based WGS methods characterize pathogens' AMR patterns and strain-types (Nadon et al., 2017). Further, WGS facilitates genomic epidemiology to identify infection origin and to track the disease transmission by location and time. Using the WGS data, genomic epidemiology predicts and tracks ongoing outbreaks, makes policy decisions to control and prevent outbreaks (Tang and Gardy, 2014; Traynor, 2009). Microbial pathogens are isolated from various sources including clinical samples, raw and processed food products and environmental samples, and cultured to isolate DNA needed for WGS. For easy-to-grow bacterial pathogens, WGS is used in nationwide outbreak monitoring systems like FoodNet (Holmes et al., 2015; Jackson et al., 2016; Joensen et al., 2014; Tack et al., 2020). However, culture based WGS methods are time consuming and labor intensive to adopt for large scale screening required for food control efforts (Nadon et al., 2017).

Whole genome enrichment (WGE) before sequencing enables WGS data generation directly from food, environmental and clinical samples without the need to isolate and culture pathogens. PCR based enrichment for sequencing of smaller viral genomes are used for genomic epidemiology and for food screening. Hybridization capture methods are used to enrich bacterial genomes using whole genome capture baits. Due to large genome sizes compared to bacteria, enrichment methods are not attempted for fungal and parasitic protist pathogens. The synthesis cost of probes to target the entire genome prohibits wider adoption of culture independent direct WGS for food screening.

Toxoplasma gondii is an intracellular parasite that causes toxoplasmosis which can manifest into severe diseases in newborns and in immunocompromised individuals (Robert-Gangneux and Dardé, 2012). *T. gondii* sexually reproduce to form oocysts only in the definitive hosts for the parasite, the felids. Felids defecate oocysts that can persist in the environment for months to years (Robert-Gangneux and Dardé, 2012; Torrey and Yolken, 2013). *T. gondii* can infect intermediate hosts that consume water or food contaminated with oocysts. *T. gondii* can spread to humans by consuming uncooked or undercooked leafy vegetables, meat and seafood (Robert-Gangneux and Dardé, 2012; Shapiro et al., 2019). Detection of oocysts in the environment is vital to prevent *T. gondii* infections in animals and humans (Shapiro et al., 2019). Detection by amplification of *T. gondii* DNA has been widely used in various matrices (DeMone et al., 2021; Kim et al., 2021; Lalonde and Gajadhar, 2016). However, PCR-based detection might fail due to inhibition or degradation of primer annealing regions. Scant parasite DNA present in environmental samples further challenge DNA amplification-based assays. To overcome these limitations, we developed a *T. gondii* detection approach using targeted DNA sequencing by whole genome enrichment (WGE).

In the previous section, we demonstrated the CNER method for WGE of difficult-to-grow bacterial pathogens to identify AMR pattern and for genomic epidemiology. Here, we demonstrate the utility of the CNER method to make WGE baits for large parasite genomes (~68 Mb) by developing WGE-CNERs for *T. gondii*. We investigate the ability of WGE-CNERs to capture different strains of *T. gondii*. We also test the sensitivity of the WGE-CNERs using different amounts of oocysts spiked into oyster lymph and test the specificity of the baits using DNA from related parasitic species.

Materials and Methods

DNA samples

We grew *T. gondii* types (type I - RH strain, type II - ME49 strain, type III - CTG strain and type X - Bobcat strain), *Sarcocystis neurona* and *Neospora hughesi* on African green monkey kidney (Vero) cell lines and collected the tachyzoites in the culture supernatant. We isolated DNA from the culture supernatant for all except *T. gondii* ME49 strain using Qiagen DNeasy Blood and Tissue Kit (Qiagen, CA, USA) as previously described (Shapiro et al., 2019). For the ME49 strain, we filtered the culture supernatant by passing through x filter to remove any remaining host cells from the tachyzoites and then isolated the DNA.

NGS library preparation

We prepared NGS libraries for the four *T. gondii* types (type I - RH strain, type II - ME49 strain, type III - CTG strain and type X - Bobcat strain,), *Sarcocystis neurona* (Sn) and *Neospora hughesi* (Nh) DNA isolated from the culture supernatants by following the Santa Cruz Reaction method for single strand NGS library preparation (Kapp et al., 2021).

For the sensitivity test, we prepared NGS libraries for the DNA isolated from the oocysts spiked in oyster hemolymph as described in (DeMone et al., 2020) using the NEB Ultra II FS kit by following manufacturer’s protocol.

Toxoplasma CNERs generation

For *T. gondii* WGE-CNERs generation, we fragmented 1 μg of culture-filtered *T. gondii* ME49 tachyzoites DNA with 0.02U DNase I at 15°C for 15 min and denatured at 95°C for 5 min. We ligated 100 ng of fragmented, denatured gDNA mixed with 35 pmol of bridge/splint oligos using T4 DNA ligase and amplified the circularized gDNA fragments as described in Chapter 2.1 and digested the RCA products using 50U of HindIII enzyme.

WGE and sequencing

For the proof-of-concept experiment we hybridized 125 ng of two replicates pooled for each of six Apicomplexan libraries with 50 ng of *T. gondii* WGE-CNERs overnight at 65°C without Human Cot-1 DNA. In a second experiment, we pooled 25 ng of three replicates for each Apicomplexan in one reaction and hybridized the pool with 100 ng of *T. gondii* WGE-CNERs overnight at 65°C with 2.5 μg Human c0t1 DNA. In both capture experiments, we enriched the captured libraries on streptavidin beads as described in Chapter 2.1 and amplified the post-capture libraries with 2X Kapa HiFi PCR mix for 20 cycles.

For sensitivity test, we pooled 75 ng each of the four libraries for oocysts spiked in oyster hemolymph samples into two pools for each biological replicates. We hybridized the pools with 75 ng of *T. gondii* WGE-CNERs overnight at 65°C with 2.5 **μ**g Human cot-1 DNA.

We amplified the post-capture libraries for 17 cycles. We purified the post-enrichment libraries using SPRI beads with 0.9X ratio, pooled the libraries in equimolar ratio and sequenced in the Illumina NextSeq with a PE 2x150 kit for ~1–3M raw read pairs.

Data analysis

We used cutadapt to remove adapter sequences and mapped the reads to the latest genome assembly of *T. gondii* ME49 strain (GCA_019455585.1, ASM1945558v1, (Xia et al., 2021) using bwa mem (Li and Durbin, 2009) with minimal exact match length set at 25 nt (-k option). ASM1945558v1 genome assembly rectified karyotype and many structural errors (Xia et al., 2021) that were present in the old reference genome. We used samtools rmdup to remove duplicate reads and filtered the mapped reads for mapping quality of 20 (-q 20 option). We counted properly paired mapped reads after removing the duplicates to determine the percent unique mapped reads.

We subsampled 50,000 raw reads, converted to fasta format and searched against the non-redundant nucleotide (nt) NCBI database (v5) using locally installed blastn. We filtered the blast results with an e-value cut-off of 1e-10. We analyzed the blastn results using the MEGAN (Huson et al., 2007) desktop tool to assign taxonomy for the reads and plotted the results.

Results

We isolated DNA from culture filtered TgME49 tachyzoites grown on Vero cell lines. To check the purity of the TgME49 DNA, we sequenced the NGS prepared using this gDNA. We mapped the reads to *T. gondii* ME49 genome assembly (ASM1945558v1) that resulted in ~72% uniquely mapped reads. We used this gDNA as *T. gondii* template DNA and prepared the WGE-CNERs as described in Chapter 2.1. The CNERs method generated 1,354 ng of CNER baits on average for four reactions using 100 ng of template DNA.

T. gondii CNERs specifically enrich *Toxoplasma gondii* DNA

We sequenced libraries before enrichment for four Tg strains and for Sn and Nh samples and mapped the reads to ASM1945558v1 genome assembly. Five technical replicates on average resulted in unique mapped reads before enrichment in the proportions of 72.2% for TgME49, 9.0% for TgRH, 10.6% for TgCTG, 15.0% for TgBobcat, 0.3% for Nh, and 0.8% for Sn (Figure 2.2.1A). In the first capture experiment, we enriched the libraries using *T. gondii* WGE-CNERs without Human Cot-1 DNA. For the TgME49 sample, unique mapped reads decreased to 62.2% due to increased duplication rate from library enrichment. For other samples, enrichment increased unique mapped read proportions in all samples: 12.6% for for TgRH, 9.9% for for TgCTG, 15.0% for TgBobcat, 0.6% fir Nh and 1.1% fo Sn. Enrichments in the presence of Human Cot-1 DNA in experiment 2 produced even more uniquely mapped reads: 80.0% for TgME49, 27.3% for TgRH, 15.0% for TgCTG, 27.8% for TgBobcat, 0.8% for Nh and 6.4% fo Sn (Figure 2.2.1A).

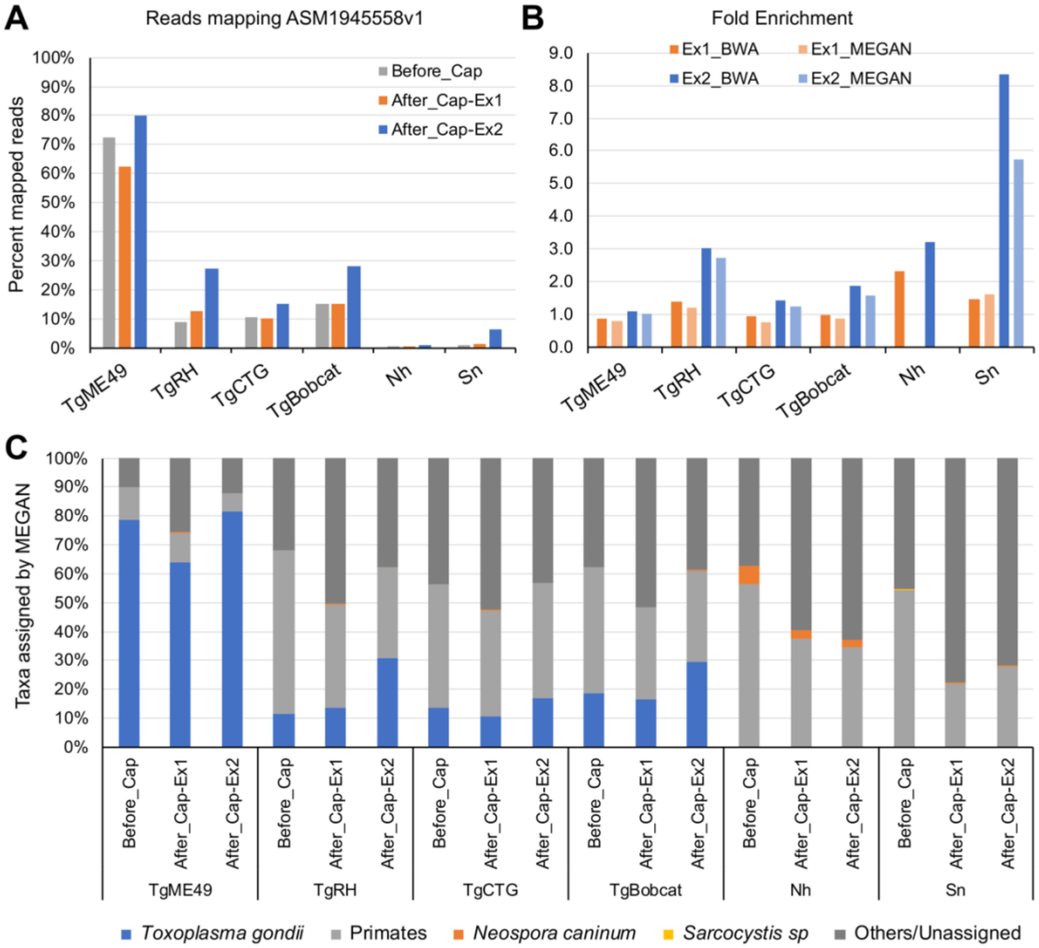


Figure 2.2.2.1 T. gondii CNERs specifically enrich Toxoplasma gondii.

(A) *Percent unique mapped reads to the T. gondii genome assembly ASM1945558v1 before (grey bars) enrichment, after enrichment in experiment 1 (orange bars) and in experiment 2 (blue bars) for four T. gondii strains (TgME49, TgRH, TgCTG, TgBobcat), N. hughesi (Nh) and S. neurona (Sn).* ***(B)*** *Fold-enrichment of T. gondii reads determined using BWA (bright colors) and MEGAN (light colors) analyses.* ***(C)*** *Percent taxa identified by MEGAN analysis of blastn results show enrichment of T. gondii reads (blue) and decrease in reads assigned to Primates (light grey).*

We calculated the enrichment efficiency as the ratio between after- and before-enrichment percent mapped reads (Figure 1B). The CNERs enrichment resulted in 0.9-fold-enrichment in experiment 1 and 1.1 in experiment 2 for TgME49, 1.4 and 3.0 for TgRH, 0.9 and 1.4 for TgCTG, 1.0 and 1.9 for TgBobcat, 2.3 and 3.2 for Nh and, 1.5 and 8.3-fold-enrichment for Sn samples (Figure 2.2.1B). Addition of Human Cot-1 DNA during hybridization in experiment 2 improved the mapped reads on average 50% for the four *T. gondii* samples compared to enrichment without it in experiment 1.

To independently verify the percent *T. gondii* reads detected before- and after-enrichment, we used the MEGAN tool for taxonomic assignment of blastn results. Due to high sequence similarity, sequence reads from the African green monkey (Vero) cell line were randomly assigned to many taxa in the Primates order including humans and chimpanzees, therefore we grouped all of these into one group as 'Primates' (Figure 2.2.1C). For DNA samples isolated from culture supernatant without filtration, 56.7%, 43.0%, 43.5% reads are assigned to Primates for the three *T. gondii* strains (TgRH, TgCTG and TgBobcat) and 56.5% for Nh and 54.2% for Sn sample before enrichments, showing a similar amounts of primate DNA in all the samples. In the TgME49 before-enrichment sequence data, we found only 11.3% reads assigned to Primates demonstrating the efficiency of culture filtering to remove host cells. Blast search assigned 78.7% of reads in TgME49 sample to T. gondii, 11.3% in TgRH, 13.5% in TgCTG and 18.8% in TgBobcat up to species level (Figure 2.2.1C) but did not identify the strain-types. We used a small subset of reads for blastn search that might not be enough to identify the strain-types. For the Nh sample, 6.2% reads identified as *Neospora* and for the Sn sample, 0.1% reads assigned to *Sarcosystis*, but no reads were identified as *T. gondii* in both samples.

MEGAN analyses of the first capture experiment without Human Cot-1 DNA identified a mixed trend in the reads assigned as T. gondii. The number of reads assigned decreased for TgME49 (64.1%) and for TgCTG (10.5%) compared to before-enrichment and increased for TgRH (13.7%) and for TgBobcat (16.4%). In the second experiment, we pooled all six samples together and added human Cot-1 DNA during the hybridization reaction. Human Cot-1 DNA is a representation of repeat-rich regions of the human genome and is used to reduce non-specific binding during DNA hybridization. Addition of human Cot-1 DNA increased the percentage of reads assigned to *T. gondii* to 81.5%, 30.7%, 16.8% and 29.5% for the four *T. gondii* samples (Figure 2.2.1C), on average a 30% increase compared to before-enrichment and 51% increase compared to hybridization without human Cot-1 DNA. Further, addition of Human Cot-1 DNA depleted reads from the host-cells resulting in 30% reduction of percentage of reads assigned as Primates (Figure 2.2.1C). Further, enrichment did not significantly change the percent reads identified as *Sarcocystis* in Sn sample and as *Neospora* in the Nh sample, indicating that the *T. gondii* WGE-CNERs does not enrich other *Sarcocystidae*.

T. gondii CNERs can detect down to 50 oocysts per ml

To determine the detection limit of WGE-CNERs method, we captured libraries prepared with DNA isolated from 5, 50, 250 and 1000 *T. gondii* ME49 oocysts spiked in per ml of hemolymph. All spike-in samples on average for two biological replicates produced few reads mapping to TgME49 reference assembly; the average was 0.5%

total and 0.2% unique mapped reads. We subsampled the after-enrichment data to determine the minimum number of reads needed to detect T. gondii. After-enrichments produced total mapped read proportions of 5.7% for 5 oocysts spike-in, 18.7% for 50 oocysts, 18.0% for 250 oocysts, and 23.4% for 1000 oocysts. These proportions remained constant under different sequencing depths (Figure 2.2.2A) and as expected, the ratio of unique reads is proportional to sequencing depth (Figure 2.2.2B).

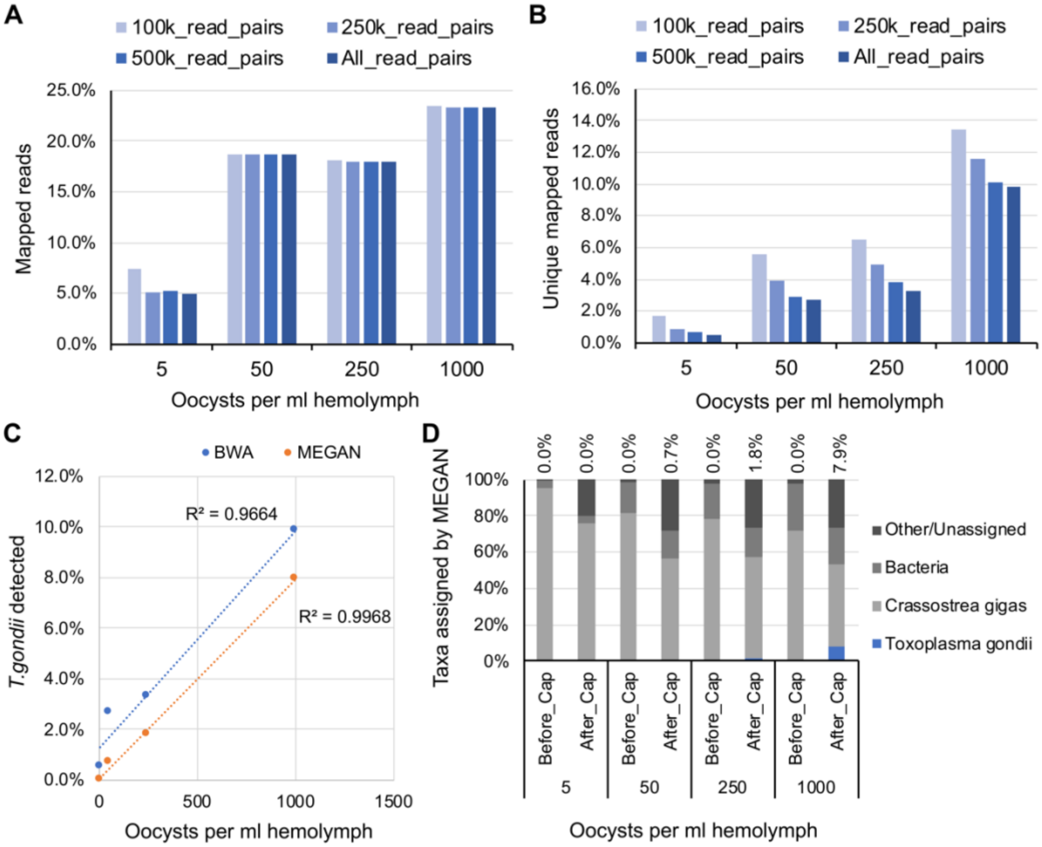


Figure 2.2.2 T. gondii CNERs can detect down to 50 oocysts per ml.

After enrichment percent total mapped reads ***(A)*** *and percent unique mapped reads* ***(B)*** *for the four spiked in hemolymph samples at the indicated number of oocysts. Shades of blue bars denote the indicated number of raw reads mapped to the T. gondii genome assembly ASM1945558v1. Scatter plot of percent T. gondii detected using*

BWA (blue) and MEGAN (orange) analyses over the number of spiked in oocysts ***(C).*** *Increasing percent of T. gondii reads (blue) identified after-enrichment that decreased the percent of oyster reads (light grey) identified by MEGAN analysis* ***(D).***

Unique mapped reads changed based on the read depth due to increased duplication rate. At 100k read pairs, 1.7% for 5, 5.5% for 50, 6.5% for 150 and 13.4% for 1000 oocysts (Figure 2.2.2B) unique reads mapped to *T. gondii* representing 7.5, 26.1, 29.9 and 62.1-fold enrichments. With all the sequenced reads (~700k read pairs), after-enrichments produced 0.5%, 2.7%, 3.3% and 9.8% unique mapped reads (Figure 2.2.2B), representing 2.2, 12.6, 15.2 and 45.4-fold enrichment (Figure 2.2.2B) for the 5, 50, 250 and 1000 oocysts per ml samples. After-enrichment with WGE-CNERs, the percent of total and unique mapped reads increased proportional to the oocysts numbers at all subsampled read counts and is highly correlated with the initial oocyst counts in the samples (Figure 2.2.2C).

We performed blastn and MEGAN analyses to validate the dose-dependent increase in the percent unique mapped reads after WGE-CNERs enrichment. In the before-enrichment data, no *T. gondii* or other *Sarcocystidae* reads identified except <20 reads assigned in the 1000 oocysts per ml sample. MEGAN assigned 95.2% of reads to Oyster for 5 oocysts/ml, 81.6% for 50 oocysts/ml, 78.6% for 250 oocysts/ml and 71.6% for 1000 oocysts/ml (Figure 2.2.2D) and 4.8% reads to other taxa (mostly bacteria associated with oyster and marine environment) for 5 oocysts/ml, 18.4% for 50 oocysts/ml, 21.4% for 250 oocysts/ml and 28.4% for 1000 oocysts/ml in the before-enrichment data (Figure 2.2.2D). After enrichment with WGE-CNERs, 0% reads assigned to *T. gondii* in the 5 oocysts/ml sample, 0.7% in the 50 oocysts/ml, 1.8% in the 250 oocysts/ml and 7.9% in the 1000 oocysts/ml samples (Figure 2.2.2C), which is highly correlated with the initial oocyst counts (Figure 2.2.2C). We achieved ~277-fold enrichment for the 1000 oocysts per ml sample. We did not determine the fold-enrichment for other concentrations due to 0% of reads identified as T. gondii. Percentage of reads assigned to Oyster decreased 30% on average to 75.7% in 5 oocysts/ml, 55.9% in 50 oocysts/ml, 55.4% in 250 oocysts/ml and 45.6% in 1000 oocysts/ml. The percentage of unassigned reads increased ~16-fold on average to 25.3% (Figure 2.2.2D) in the after-enrichment data.

Discussion

Our WGE-CNERs method detected as few as 50 oocysts per ml with as low as 100k raw read pairs confirmed by two bioinformatic analyses. We detected *T. gondii* reads at the lowest tested concentration of 5 oocysts per ml in one of our analyses. We spiked the oocysts in oyster hemolymph to mimic the real-world shellfish food samples. Further studies are needed to test the performance of WGE-CNERs method to detect *T. gondii* from other food and environmental matrices, and to determine the lowest sequencing reads required to detect the lowest number of oocysts.

Current DNA amplification based *T. gondii* detection methods including conventional, nested and quantitative PCR assays detect up to 1 - 5 oocysts (Kim et al., 2021; Lalonde and Gajadhar, 2016; Sotiriadou and Karanis, 2008; Villena et al., 2004). The 18S rRNA metabarcode sequencing was demonstrated to detect down to 5 oocysts (DeMone et al., 2020). Though these methods achieve the lowest detection thresholds, they suffer a poor reproducibility rate (0 - 100%). The inconsistency was due to PCR inhibition (Shapiro et al., 2019; Sotiriadou and Karanis, 2008; Villena et al., 2004) and degradation of priming sites (Nichols et al., 2018). DNA isolated from environmental matrices is usually degraded and often fails to amplify in any PCR based methods.

The hybridization capture method using WGE-CNERs would be an efficient alternative for PCR amplification-based methods to detect *T. gondii* from diverse DNA samples. Further, the WGE-CNERs method enriches the entire genome of T. gondii, unlike the target region amplification and 18s rRNA metabarcoding methods. The WGS data generated using the WGE-CNERs can be used for genome wide SNP genotyping for strain identification (Su et al., 2012) and to determine the copy-number variation in the virulent genes to predict the pathogenicity (Lorenzi et al., 2016). For these advantages, the WGE-CNERs method would be an efficient tool for genomic epidemiology and pathogen surveillance of any food and waterborne pathogens.

Pathogen surveillance programs are needed for swift outbreak identification and environmental monitoring to control the transmission of pathogens among humans, animals, wildlife and the environment to achieve One Health goals. One Health aspect of *T. gondii* infection was realized after the detection of waterborne outbreaks in Brazil (Balbino et al., 2022) and endangerment of marine mammals (Dubey et al., 2020). The WGE-CNERs can be adopted for One Health surveillance programs using various food and environmental sample matrices.

Acknowledgements

The authors thank the members of Karen Shapiro lab at UC Davis, especially Andrea Packham, Lauren Camp, Lezlie Rueda, Minji Kim and Anika Patel for providing the parasite DNA.

References

Balbino, L.S., Bernardes, J.C., Ladeia, W.A., Martins, F.D.C., Nino, B. de S.L., Mitsuka-Breganó, R., Navarro, I.T., Pinto-Ferreira, F., 2022. Epidemiological study of toxoplasmosis outbreaks in Brazil. Transbound. Emerg. Dis. 69, 2021–2028. doi:10.1111/tbed.14214

Brown, E., Dessai, U., McGarry, S., Gerner-Smidt, P., 2019. Use of whole-genome sequencing for food safety and public health in the United States. Foodborne Pathog. Dis. 16, 441–450. doi:10.1089/fpd.2019.2662

DeMone, C., Hwang, M.-H., Feng, Z., McClure, J.T., Greenwood, S.J., Fung, R., Kim, M., Weese, J.S., Shapiro, K., 2020. Application of next generation sequencing for detection of protozoan pathogens in shellfish. Food Waterborne Parasitol. 21, e00096. doi:10.1016/j.fawpar.2020.e00096

DeMone, C., Trenton McClure, J., Greenwood, S.J., Fung, R., Hwang, M.-H., Feng, Z., Shapiro, K., 2021. A metabarcoding approach for detecting protozoan pathogens in wild oysters from Prince Edward Island, Canada. Int. J. Food Microbiol. 360, 109315. doi:10.1016/j.ijfoodmicro.2021.109315

Dubey, J.P., Murata, F.H.A., Cerqueira-Cézar, C.K., Kwok, O.C.H., Grigg, M.E., 2020. Recent epidemiologic and clinical importance of *Toxoplasma gondii* infections in marine mammals: 2009-2020. Vet. Parasitol. 288, 109296. doi:10.1016/j.vetpar.2020.109296

Holmes, A., Allison, L., Ward, M., Dallman, T.J., Clark, R., Fawkes, A., Murphy, L., Hanson, M., 2015. Utility of Whole-Genome Sequencing of Escherichia coli O157 for Outbreak Detection and Epidemiological Surveillance. J. Clin. Microbiol. 53, 3565–3573. doi:10.1128/JCM.01066-15

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res. 17, 377–386. doi:10.1101/gr.5969107

Jackson, B.R., Tarr, C., Strain, E., Jackson, K.A., Conrad, A., Carleton, H., Katz, L.S., Stroika, S., Gould, L.H., Mody, R.K., Silk, B.J., Beal, J., Chen, Y., Timme, R., Doyle, M., Fields, A., Wise, M., Tillman, G., Defibaugh-Chavez, S., Kucerova, Z., Gerner-Smidt, P., 2016. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin. Infect. Dis. 63, 380–386. doi:10.1093/cid/ciw242

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., Aarestrup, F.M., 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J. Clin. Microbiol. 52, 1501–1510. doi:10.1128/JCM.03617-13

Kapp, J.D., Green, R.E., Shapiro, B., 2021. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J. Hered. 112, 241–249. doi:10.1093/jhered/esab012

Kim, M., Shapiro, K., Rajal, V.B., Packham, A., Aguilar, B., Rueda, L., Wuertz, S., 2021. Quantification of viable protozoan parasites on leafy greens using molecular methods. Food Microbiol. 99, 103816. doi:10.1016/j.fm.2021.103816

Kirk, M.D., Pires, S.M., Black, R.E., Caipo, M., Crump, J.A., Devleesschauwer, B., Döpfer, D., Fazil, A., Fischer-Walker, C.L., Hald, T., Hall, A.J., Keddy, K.H., Lake, R.J., Lanata, C.F., Torgerson, P.R., Havelaar, A.H., Angulo, F.J., 2015. World health organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: A data synthesis. PLoS Med. 12, e1001921. doi:10.1371/journal.pmed.1001921

Lalonde, L.F., Gajadhar, A.A., 2016. Detection of Cyclospora cayetanensis, Cryptosporidium spp., and *Toxoplasma gondii* on imported leafy green vegetables in Canadian survey. Food and Waterborne Parasitology 2, 8–14. doi:10.1016/j.fawpar.2016.01.001

Law, J.W.-F., Ab Mutalib, N.-S., Chan, K.-G., Lee, L.-H., 2014. Rapid methods for the detection of foodborne bacterial pathogens: principles, applications, advantages and limitations. Front. Microbiol. 5, 770. doi:10.3389/fmicb.2014.00770

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Lorenzi, H., Khan, A., Behnke, M.S., Namasivayam, S., Swapna, L.S., Hadjithomas, M., Karamycheva, S., Pinney, D., Brunk, B.P., Ajioka, J.W., Ajzenberg, D., Boothroyd, J.C., Boyle, J.P., Dardé, M.L., Diaz-Miranda, M.A., Dubey, J.P., Fritz, H.M., Gennari, S.M., Gregory, B.D., Kim, K., Sibley, L.D., 2016. Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic *Toxoplasma gondii* genomes. Nat. Commun. 7, 10147. doi:10.1038/ncomms10147

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., Gilpin, B., Smith, A.M., Man Kam, K., Perez, E., Trees, E., Kubota, K., Takkinen, J., Nielsen, E.M., Carleton, H., FWD-NEXT Expert Panel, 2017. PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill. 22. doi:10.2807/1560-7917.ES.2017.22.23.30544

Nichols, R.V., Vollmers, C., Newsom, L.A., Wang, Y., Heintzman, P.D., Leighton, M., Green, R.E., Shapiro, B., 2018. Minimizing polymerase biases in metabarcoding. Mol. Ecol. Resour. doi:10.1111/1755-0998.12895

Ray, L.C., Griffin, P.M., Wymore, K., Wilson, E., Hurd, S., LaClair, B., Wozny, S., Eikmeier, D., Nicholson, C., Burzlaff, K., Hatch, J., Fankhauser, M., Kubota, K., Huang, J.Y., Geissler, A., Payne, D.C., Tack, D.M., 2022. Changing Diagnostic Testing Practices for Foodborne Pathogens, Foodborne Diseases Active Surveillance Network, 2012-2019. Open Forum Infect. Dis. 9, ofac344. doi:10.1093/ofid/ofac344

Robert-Gangneux, F., Dardé, M.-L., 2012. Epidemiology of and diagnostic strategies for toxoplasmosis. Clin. Microbiol. Rev. 25, 264–296. doi:10.1128/CMR.05013-11

Shapiro, K., Bahia-Oliveira, L., Dixon, B., Dumètre, A., de Wit, L.A., VanWormer, E., Villena, I., 2019. Environmental transmission of Toxoplasma gondii: Oocysts in water, soil and food. Food and Waterborne Parasitology 15, e00049. doi:10.1016/j.fawpar.2019.e00049

Sotiriadou, I., Karanis, P., 2008. Evaluation of loop-mediated isothermal amplification for detection of *Toxoplasma gondii* in water samples and comparative findings by polymerase chain reaction and immunofluorescence test (IFT). Diagn. Microbiol. Infect. Dis. 62, 357–365. doi:10.1016/j.diagmicrobio.2008.07.009

Su, C., Khan, A., Zhou, P., Majumdar, D., Ajzenberg, D., Dardé, M.-L., Zhu, X.-Q., Ajioka, J.W., Rosenthal, B.M., Dubey, J.P., Sibley, L.D., 2012. Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. Proc Natl Acad Sci USA 109, 5844–5849. doi:10.1073/pnas.1203190109

Tack, D.M., Ray, L., Griffin, P.M., Cieslak, P.R., Dunn, J., Rissman, T., Jervis, R., Lathrop, S., Muse, A., Duwell, M., Smith, K., Tobin-D’Angelo, M., Vugia, D.J., Zablotsky Kufel, J., Wolpert, B.J., Tauxe, R., Payne, D.C., 2020. Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2016-2019. MMWR Morb Mortal Wkly Rep 69, 509–514. doi:10.15585/mmwr.mm6917a1

Tang, P., Gardy, J.L., 2014. Stopping outbreaks with real-time genomic epidemiology. Genome Med. 6, 104. doi:10.1186/s13073-014-0104-4

Torrey, E.F., Yolken, R.H., 2013. Toxoplasma oocysts as a public health problem. Trends Parasitol. 29, 380–384. doi:10.1016/j.pt.2013.06.001

Traynor, B.J., 2009. The era of genomic epidemiology. Neuroepidemiology 33, 276–279. doi:10.1159/000235639

Villena, I., Aubert, D., Gomis, P., Ferté, H., Inglard, J.-C., Denis-Bisiaux, H., Dondon, J.-M., Pisano, E., Ortis, N., Pinon, J.-M., 2004. Evaluation of a strategy for *Toxoplasma gondii* oocyst detection in water. Appl. Environ. Microbiol. 70, 4035–4039. doi:10.1128/AEM.70.7.4035-4039.2004

Xia, J., Venkat, A., Bainbridge, R.E., Reese, M.L., Le Roch, K.G., Ay, F., Boyle, J.P., 2021. Third-generation sequencing revises the molecular karyotype for *Toxoplasma gondii* and identifies emerging copy number variants in sexual recombinants. Genome Res. 31, 834–851. doi:10.1101/gr.262816.120

Chapter 3. Targeted genotyping of DNA isolated from rootless hair

Abstract

Forensic or Investigative genetic genealogy (F/IGG) requires genotyping of SNP markers. Current genotyping methods fail to generate profiles due to sparse, degraded and contaminated DNA isolated from rootless hairs. Genotyping by whole genome sequencing overcomes some of these disadvantages but is costly. Targeted sequencing of SNPs is a cost-effective method for genotyping by sequencing. However, capture baits for thousands of F/IGG informative SNP markers does not exist and is costly to make, therefore prohibits wider adoption of targeted sequencing for F/IGG. We developed the circular nucleic acid enrichment reagent (CNER) synthesis method to generate microgram quantities of capture baits for targeted sequencing. We selected ~108k SNPs from the union of SNPs in three DTC platforms and designed three 12k panels and a 72k panel. I tested hybridization capture protocols using NGS libraries made with NA12878 DNA and DNA isolated from rootless hair samples from a volunteer. The preliminary results show that the GC content of the target regions affect the capture efficiency. I am performing captures on the remaining 50 volunteers. We will test the concordance between genotypes generated using low-coverage WGS and targeted sequencing. We will test the cost-benefit of targeted genotyping to generate genotype profiles for F/IGG.

Introduction

Individuals voluntarily deposit their genetic information in public databases like GEDmatch driven by their curiosity to search relatives and to learn about their genealogy. The GEDmatch database contains about 1.4 million genetic profiles due to exponential growth of direct-to-consumer (DTC) genetic testing [1, 2]. The DTC tests use microarrays to genotype ~650k SNP makers across the human genome [2, 3]. The number of individual profiles available with high dense SNP markers facilitate genetic genealogy search to find distant relatives [2, 4]. In 2018, law enforcement agencies in collaboration with private genealogists searched the database to solve decades old Golden State Killer case [5]. Success of solving this cold case using the DTC database search led to the beginning of the forensic or investigative genetic genealogy (F/IGG) field [1, 6].

F/IGG search requires genotyping trace DNA obtained from biological specimens and touched objects [1, 2]. However, trace DNA is scarce, degraded and contaminated with unwanted DNA due to the source and storage of the forensic specimen [1, 7]. DNA from hair is fragmented [8] due to the action of specific endonucleases during the keratinization process [9]. DNA from human remains are degraded and contaminated due to environmental exposure [10, 11]. DNA of decades-old specimens might also be degraded due to poor storage [12, 13]. Microbial DNA constitutes the majority of unwanted DNA in many forensic samples, confounding forensic analysis [10]. Conventional genotyping methods like single base extension and microarray are inadequate for F/IGG due to their technical limitations for analyzing trace DNA [2]. Single base extension methods are low-throughput and suffer from PCR stutter and electropherogram artifacts [14]. Though microarray methods can genotype thousands of SNPs [15, 16], the high quality and quantity of input DNA requirement impede their use for F/IGG genotyping [2, 17, 18].

Genotyping by sequencing (GBS) using next generation sequencing (NGS) methods is currently the widely used method for F/IGG analyses of trace DNA [2] (Green et al 2023). GBS methods use whole genome sequencing (WGS) and targeted sequencing approaches. In the WGS approach, mid-to-high coverage whole genome data is generated from trace DNA by extensive sequencing [19, 20]. The entire DNA sample from the specimen is converted into a sequenceable library by ligating universal adapters [19–21] (Green et al 2023) that reduce the cost-effectiveness by inadvertent sequencing of unwanted DNA [22]. Imputation methods are then used to make statistical predictions of the genotype of the specimen using allele frequency information from the 1000 genome project [2].

Targeted sequencing of SNP loci is cost-effective to produce higher coverage than WGS that enables direct genotyping and eliminates statistical imputation. Targeted genotyping is performed by PCR enrichment and by hybridization capture enrichment methods. PCR enrichment using target SNP-specific primers is used for mixture resolution [23], identification of missing persons [24], human remains identification [25, 26], and kinship analysis and paternity tests [27]. However, PCR enrichment is limited to a few hundred SNP markers that is inadequate for F/IGG analyses. Further, PCR enrichment of degraded samples results in allelic dropouts due to the loss of priming sites [28]. Contaminants like melanin and collagen from DNA isolation cause PCR inhibition to further aggravate allelic dropout [29–31].

Hybridization capture methods use complementary DNA or RNA probes specific to target SNP regions for targeted genotyping of a few hundred forensically relevant SNPs [32–34]. Recently SNP hybridization capture panels were designed for F/IGG applications [35, 36] and for kinship analysis [37]. The FORCE panel targets 5,422 SNPs designed to identify ancestry [36]. The Kintelligence panel targets 10,230 SNPs to identify kinship [35]. Gorden reported two panels with 25K and 95K SNP markers to analyze kinship using degraded bone samples [37]. Though these panels improve the number of target SNPs analyzed for F/IGG, they still fall short on the number of SNPs analyzed by microarray due to the cost to make and validate large panels of SNP enrichment reagents.

We developed the CNER method to cost-effectively make hybridization capture probes for a large number of SNP markers for targeted genotyping (Chapter 1). Here, we demonstrate the CNER method to generate a capture panel to enrich 108k SNPs common to three major DTC platforms. From this large panel, we selected 36k SNPs for enrichment from NGS libraries made using rootless hair samples collected from 50 volunteers. We show high concordance between the genotypes generated using the CNERs enrichment and using WGS data of the same hair samples and microarray genotypes generated using the spit DNA. We expect our large collection of DTC SNP CNERs will be a valuable tool for F/IGG searches and for other forensic applications including missing person and victim identification.

Materials and Methods

Rootless hair DNA libraries

Previously we described the collection and WGS data generation of rootless hair samples collected from 50 volunteers (Green et al, 2023). We pooled 33.3 ng each of the three library replicates made for each hair sample and re-amplified using 2x KAPA

HiFi master mix with universal Illumina amplification primers for 8 cycles. We cleaned the amplified libraries using SPRI beads at 1.2x ratio and used them for the capture experiments.

Forensic panels design

F12k_Kin

Recently Snedecor et al designed a SNP panel to analyze kinship [33]. We chose 10,148 SNPs from this panel. To fill the panel to 12k oligos, we chose proxy SNPs using the NCI LDproxy API tool [36]. We filtered the proxy sites that are more than 1 kb apart and have r2 > 0.5. For a non-overlapping set of 308 SNPs in the kintelligence panel, we picked one, two and three proxy sites, 1852 SNPs in total. We chose 81 bp target regions around the SNP loci and did not filter for any parameters (homopolymer, repeat regions and frequently occurring kmer) as this panel was previously validated. We appended the AscII site and oligo-dT linkers and synthesized the panel as an oligo pool (Twist Biosciences).

The F12k_40GC and F12k_55GC panels

We used the 2,068,959 biallelic autosomal SNP sites from the three major DTC genetics platforms [3]. We made a fasta file for the 80 bp region around the SNP loci and filtered out sites that have target regions with homopolymers (>=6 nt) and 17mers frequently occurring in the human genome. To test whether GC content of the target regions has any effect on SNP enrichment, we selected two sets of 12k SNPs each with target regions at 40% (F12k_GC40) and 55% GC content (F12k_GC55). We appended the AscII site and oligo-dT at both ends of 80 bp target regions and synthesized two panels individually as oligo pools (Twist Biosciences).

The F72k panel

We filtered the autosomal biallelic DTC SNPs based on the normalized coverage in the 50 volunteer panel hair DNA WGS data between 1 - 2x and retained 1,417,072 sites. We filtered these sites for MAF of 30 - 50 % to retain 236,577. From these, we selected 132,095 sites that did not overlap with repeat regions using the repeat masker database. For this final set of SNP sites, we chose 81 bp target region by selecting 40 nt up and downstream of the SNP loci and made a fasta file. We filtered this target region fasta file for homopolymers (>=6 nt) and 17mers frequently occurring in the human genome. We chose 70,956 sites that had 40 - 70% GC in the 81 bp target region and topped-off with 1,044 sites that have 39.5 - 40% target region GC content to design the final F72k panel. We randomly split these 72k target regions into three sets of 24k targets each, appended the AscII site and oligo-dT linker and synthesized them as three oligo pools (Twist Biosciences).

We filtered out 4,882 SNPs which we did enrich in the preliminary hair captures and in the NA1278 sample. For a final set of three panels each with 12k SNPs, we split the remaining 67,118 sites into three GC bins: 40 - 44% (F12k_1), 44 - 47% (F12k_2) and 47- 51% (F12k_3). The first two bins had 12,531 and 13,083 SNPs from which we randomly chose 12k and the third bin had 11,965 SNPs. We synthesized these three panels individually as three oligo pools (Twist Biosciences).

CNERs generation

We made CNERs for each of the forensic SNP panels by following the CNER method described in Chapter 1. Briefly, we bulk circularized the oligo pools by oligo-dA splinted ligation using T4 DNA ligase. We amplified the circularized oligo templates by rolling circle amplification (RCA) at 30°C for 24 - 40 hr using phi29 polymerase with biotinylated dATP and dUTP. We digested the high molecular weight DNA generated by RCA using AscII enzyme at 37°C for 5 hr and SPRI cleaned the monomeric CNERs.

DTC SNP enrichment using CNERs

We followed the hybridization capture methods described in Chapter 1 for the DTC SNP enrichments with the following modifications. For various DTC SNP panel validations, we made NGS libraries with the NA12878 DNA using a single strand library preparation method by following the Santa Cruz Reaction [19]. We hybridized 100 - 300 ng of libraries with 10 - 30 ng of CNERs of indicated panels for 19.5 hr at 60 - 70°C as specified. We sequenced the post-enrichment libraries in Illumina NextSeq with 2x75 cycles to generate ~1.5M raw read pairs.

For hair library captures, we pooled 20 ng each of three library replicates of two individual hair samples (S001 and S003) into a 6-plex pool with 120 ng total library and captured the pool with 24 ng of F72k CNERs panel at 65°C for 17.5hr. We amplified the post-enrichment libraries for 17 cycles with universal Illumina amplification primers using 2x KAPA HiFi master mix (Roche) and cleaned using SPRI beads (1.2x) as described in Chapter 1. We quantified the post-capture libraries using Qubit (Thermo Fisher) and in Illumina NextSeq with 2x75 cycles to generate ~2M raw read pairs per library.

Data Analyses

We removed adapter sequences and merged overlapping paired end reads using SEQPREP2 (https://github.com/jeizenga/SeqPrep2). We mapped the merged reads to the GRCh38 reference using BWA ALN - v0.7.17-r1188. We assigned zero map quality for unmapped reads using PICARD CLEANSAM - v2.21.7 and marked duplicated reads using PICARD MARKDUPLICATES - v2.21.7. We determined read coverage at target SNPs using BEDTOOLS MULTICOV - v2.29.1 and counted the coverage metrics using AWK. We merged the clean bam files for three library replicates using SAMTOOLS MERGE - 1.10 and recounted the SNP coverage metrics for a given hair sample. We plotted SNP coverage against GC content, and percent targets using custom python scripts (https://github.com/bsun210/CNERs_ancient_horses). We used SAMTOOLS MPILEUP - 1.10 with the option to output GP and GQ flags to call variants on the library replicates merged bam files for each hair sample. We used VCFTOOLS - 0.1.16 to find the concordant and discordant sites. We plotted the histogram of concordant and discordant sites based on GC content and coverage depth at the variant site using custom python scripts (https://github.com/bsun210/CNERs_Forensics). We performed a nonparametric Mann-Whitney Wilcoxon (MWW) rank test for comparison between groups.

Results

Major DTC providers test ~600k SNPs each but the most of these target SNPs does not overlap between different providers. Lu et al clustered SNPs genotyped in major DTC providers into six groups based on overlapping SNPs, platform and version history [3]. From the cluster analyses, they find that only ~23% of the total 2,135,214 SNPs from the four major DTC platforms overlap between any two cluster/platforms (Figure 3.1A). We used the union of 2,135,214 SNPs and plotted the histogram of GC contents for the 80bp target regions around the SNP loci which produced a maxima at 40% GC content (Figure 3.1B).

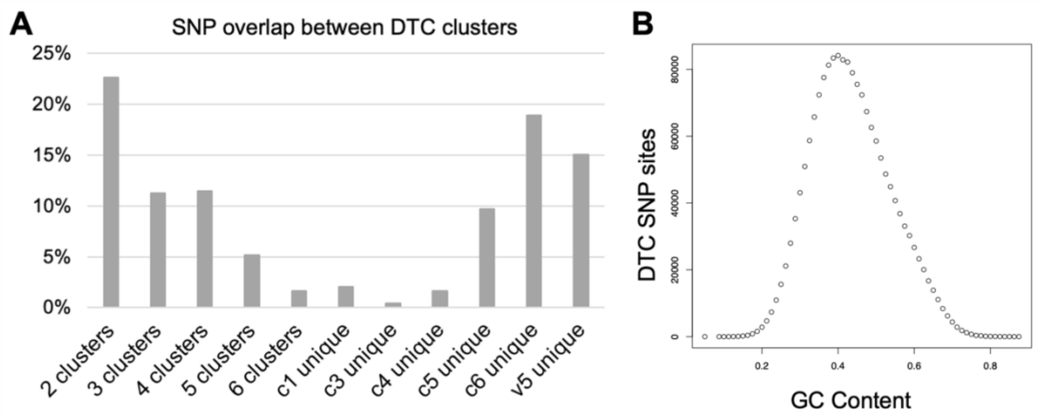


Figure 3.1 SNP maker overlap between DTC test providers.

(A) *Percentage of overlapping between clusters and percentage of SNPs unique in different DTC tests.* ***(B)*** *and t unique mapped reads to the T. gondii genome assembly ASM1945558v1*

Our previous work on targeted genotyping of ancient DNA (Chapter 1) showed that CNERs enrichment produces higher SNP coverage for target regions with ~50 - 55% GC content. To test the performance of CNERs across different GC contents, we designed two panels of 12,000 SNPs each, exclusively on target regions with 40% GC content for the F12k_GC40 panel and 55% GC content for the F12k_GC55 panel. We also designed the F12k_Kin panel to target ~10k SNPs used for kinship analyses [33] and filled the panel up to 12k targets with proxy SNP sites. The CNER method produced 377 ng of probes for the F12k_GC40 panel, 1143 ng of probes for the F12k_GC55 panel and 932 ng of probes for the F12k_Kin panels. We mixed the oligo templates for all three panels and made a combined F36k CNERs panel to target ~36k SNPs which generated 1820 ng of CNERs probes.

We captured libraries made using the NA12878 DNA to test the effect of hybridization temperature and target region GC contents on SNP captures. We defined the SNP Enrichment Efficiency as the percentage of total or unique mapped reads mapped to target SNP loci to measure the success of enrichment. The F12k_40GC panel produced 27.3% and 26.9% SNP_EE (unique) which is comparable to 24.8% and 25.5% produced by the F12k_55GC panel for hybridizations at 62°C and 65°C (Figure 3.2A). Due to low duplicate reads, the SNP_EE determined using both the total reads and unique reads are similar for the NA12878 DNA sample (Figure 3.2A and Table S1).

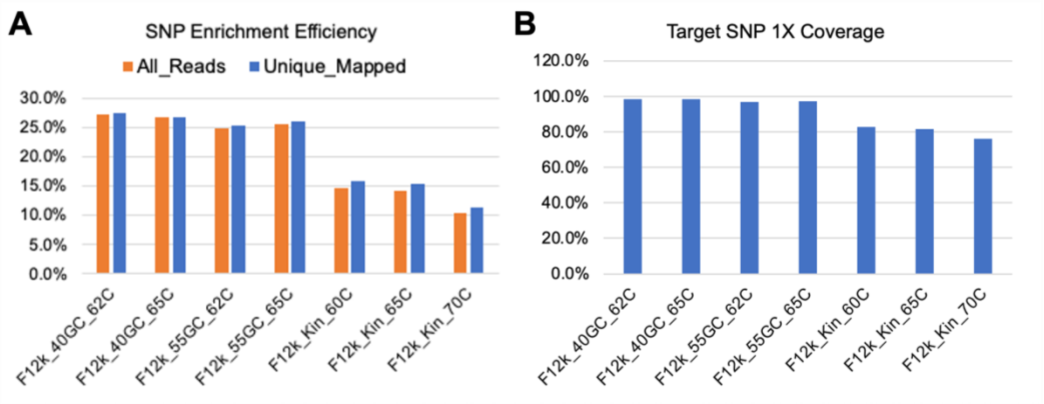


Figure 3.2 Hybridization optimization for F12k panels.

(A)*SNP enrichment efficiency determined using all reads (orange bars) and unique mapped reads (blue bars) for the three F12k panel captured at different hybridization temperatures.* ***(B)*** *Percent target SNPs covered with at least one read (1X coverage) for the three panels captured at different tempratures.*

Using 250k raw read pairs, the enrichments covered 96.7% - 98.6% of target SNPs with at least one read (1X coverage, Figure 3.2B). These results indicate that the two panels targeting different GC contents hybridized at two different temperatures produce comparable SNP enrichments when captures performed individually. We observed an identical SNP_EE (~39%) for the two GC panels in a replicate experiment hybridized at 65°C (Figure 3.3A) and recovered ~99.5% target SNPs with ~33x mean unique coverage using 1M raw read pairs (Figure 3.3B).

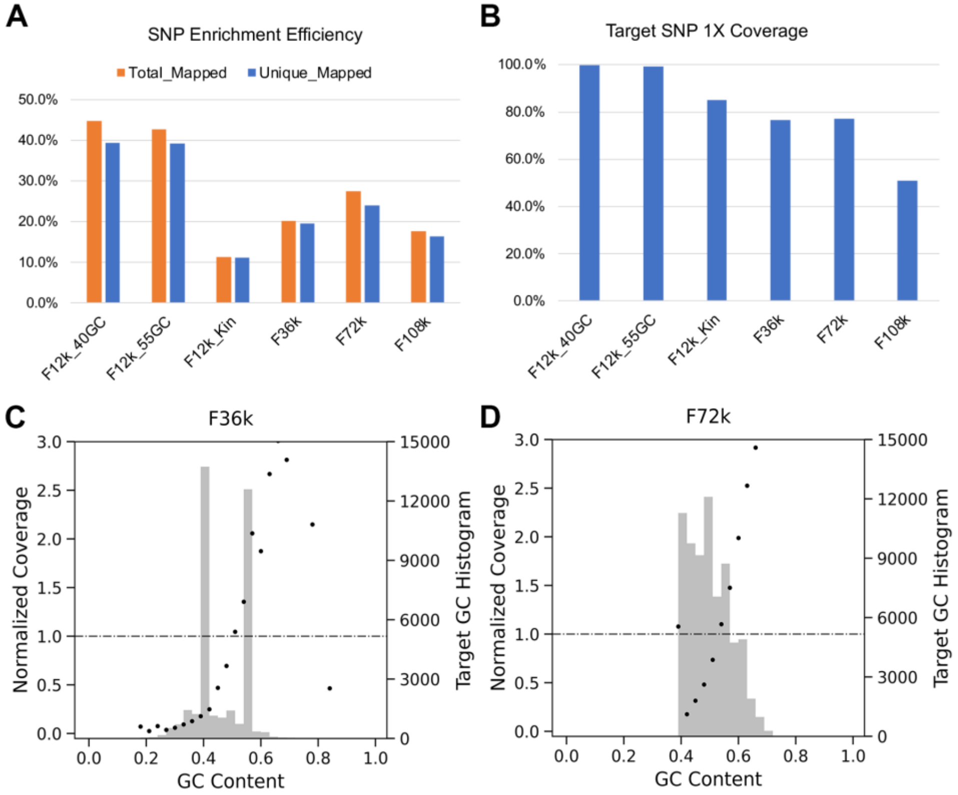


Figure 3.3 Performance of CNER panels to enrich DTC SNPs.

*SNP enrichment efficiency **(A)** and percent target SNPs covered with at least one read **(B)** for the three F12 panels captured individually, mixed together (F36k), for the F72K panel designed based on hair coverage data and all panels mixed together to target 108K SNPs (F108k). Mean of normalized coverage (primary Y-axis) plotted across GC content of CNER target regions for the F36k panel **(C)** and F72k panel **(D)** show that* *high GC (>50%) have sample-normalized coverage of 1 or higher. A histogram of GC bins across the target regions is shown in the secondary Y-axis.*

The F12k_Kin panel produced 15.9% SNP_EE at 60°C, 15.3% SNP_EE at 65°C and 11.4% SNP_EE at 70°C hybridization temperatures (Figure 3.2A) and recovered 82.7%, 81.6% and 76.1% target SNPs using 250k raw read pairs (Figure 3.2A). We sequenced a replicate for 1M raw read pairs that produced 11.2% SNP_EE and 85.2% target SNPs covered with 9X unique mean coverage. We find an increased number of proxy SNP target regions containing repetitive elements in the F12k_Kin panel which might be causing the lower enrichment efficiency compared to the GC panels. Among the 1852 proxy SNPs, 45% loci contain LINE/SINE elements compared to only 27% target regions in the Kintelligence panel and 28% in the F12k_40GC and 20% in the F12k_40GC SNPs.

The F36K panel on average produced 19.6% SNP_EE, 76.5% 1X coverage (Figure 3.3A and B) and 6.3X unique mean target SNP coverage for two replicates. We analyzed the three SNP panels from F36k capture data individually and found that the F12k_GC55 dominates the other two panels. SNPs in the F12k_GC55 produced 13.7% SNP_EE compared to 3.0% for F12k_GC40 and 3.6% for the F12k_Kin panel. Further, the percent target SNPs covered with at least one read is 97.3% for the F12k_GC55 compared to 74.6% for F12k_GC40 and 59.6% for the F12k_Kin panel (Table S1). We speculate that the hybridization kinetics change when CNERs of different GC content are mixed together. The hybridization conditions that we used favors enrichment of regions with higher GC content.

We generated WGS data from rootless hair samples collected from 50 volunteers (Green et al, unpublished data). For the ~2M DTC SNPs, we determined the average normalized coverage from the WGS data from all hair samples (Green et al). We designed a new CNERs panel to target 72k SNPs specifically to genotype DNA from rootless hair samples. For this panel, we chose SNPs that had 1 - 2x average normalized coverage in the WGS data, MAF less than 0.3 and fall within target regions with 40 - 70% GC content. We synthesized the F72k panel into three randomly selected 24k oligo pools and combined them to make the F72k CNERs panel. The F72K panel on average produced 24.0% SNP_EE, 77% 1X coverage (Figure 3.3A and B) and 5X unique mean target SNP coverage for two replicates captured using NA12878 libraries from 1M raw read pairs. The three 24k panels performed comparably when analyzed individually from the F72k data. The three panels on average produced 11.5% SNP_EE, 80% 1X coverage and 7.2X unique mean target SNP coverage (Table S1).

To test the performance of the F72k panel for hair libraries, we pooled three library replicates made from two individual hair samples (S001 and S003) into a 6-plex pool and captured the pool using the F72k panel. We aimed to sequenced 2M raw read pairs for each library. Due to pooling inconsistencies, S001 hair libraries produced on average 366k raw read pairs, therefore we did not analyze this sample further. Three library replicates of the S003 hair sample produced ~3.5M raw read pairs each which resulted in 2.7% SNP_EE (unique read). The unique reads mapping to SNP loci are low due to the unique genome copies available for sequencing in the hair libraries (Green et al). When we looked at the total reads, on average 57.8% of the total reads mapped to F72k SNP loci, indicating that the CNERs enrich reads with target SNP loci higher than background. However, due to the nature of the hair samples which are known to have lower amounts of fragment DNA with few unique genome copies, the unique read SNP_EE is lower.

We merged the reads from three library replicates for the S003 sample resulting in 10.6M raw read pairs that covered 64.5% (46,455) targeted SNPs with at least one read (1X coverage) and produced 3.5X unique mean coverage. We explored the characteristics of 25,544 target SNPs which had zero-coverage. Among the 25,544 zero-coverage SNPs, 17,510 are in target loci with 39.5 - 47% GC content which is significantly (MWW p-value < xx) different from the distribution of GC content for all 72k target SNPs (Figure 3.4A). We also find that the zero-coverage SNPs have significantly (MWW p-value 1.35e-201) low average normalized coverage in the hair WGS data (Figure 3.4B). These results show that the lower enrichment efficiency of CNERs for targets with low GC (<47%) compounded with difficulty in sequencing reads covering these regions in the rootless hair DNA library led to target SNP drop out.

Next, we explored the concordance between the genotypes from the WGS data and CNERs capture data for the same hair sample. We generated high coverage (x) WGS data using DNA isolated from the saliva samples from the same individuals who donated hair samples. We called genome-wide variants in this WGS data using the GATK pipeline (Green et al). For the CNERs enrichment data, we called genotypes using samtools mpileup. We used bcftools to find 31,550 SNPs that have the same genotypes (concordant sites) determined using both WGS and capture data, and for 14,478 SNPs the genotypes differed (discordant sites) between the two data.

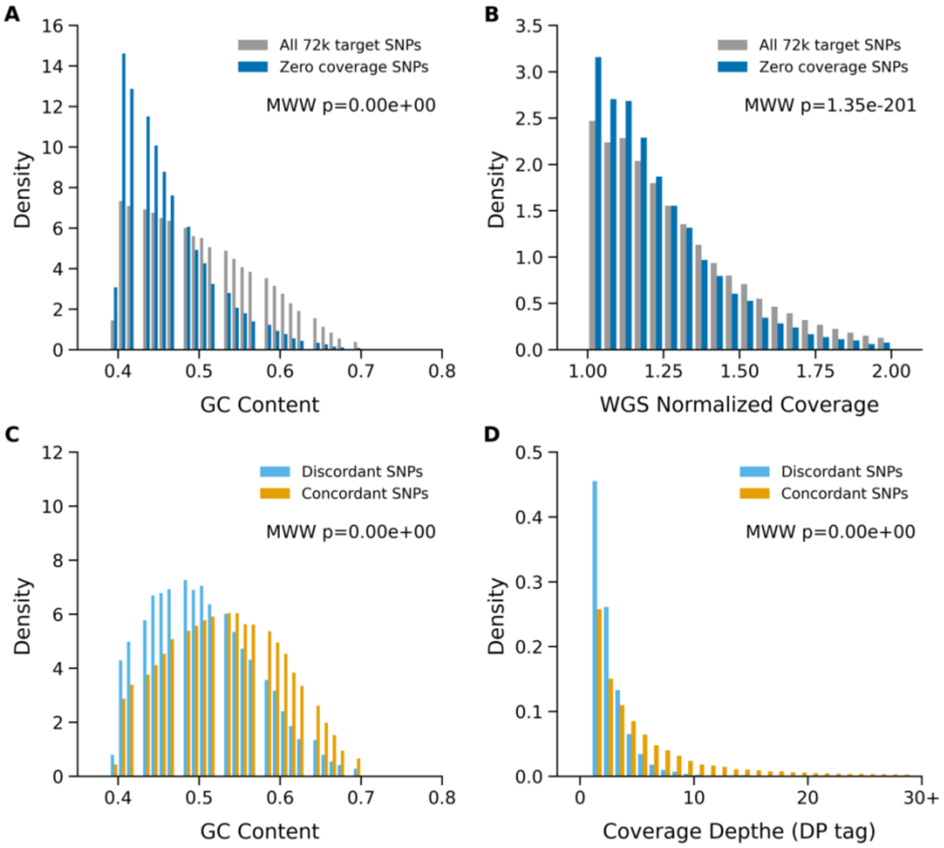


Figure 3.4 Performance of F72k CNERs panel to enrich SNPs from hair sample. *Histogram density plots of across the GC content **(A)** and normalized coverage in whole genome sequencing of hair sample **(B)**, for all target regions in the F72k CNERs panels (grey bars) and for target regions of SNPs not enriched (Zero coverage SNPs, blue bar) in the hair sample. Histogram density plots of across the GC content **(C)** and coverage depth **(D)**, for SNP genotypes concordant (orange bars) and discordant (cyan bars) between the CNERs enrichment data and the WGS data.*

We explored target region GC content and found that discordant sites are significantly (MWW p-value < xx) more prevalent in low GC regions and concordant sites are enriched in high GC regions (Figure 3.4C). We also find that the discordant sites significantly (MWW p-value < xx) have lower coverage depth compared to concordant sites (Figure 3.4D). These results show that SNPs with lower read coverage due to lower enrichment efficiency by CNERs in the low GC regions lead to disagreement between genotypes called using the capture data and WGS data. Alternatively, lower coverage in the WGS data at these sites might also lead to genotype errors that cause discordance between genotypes called from CNERs enrichment data.

We plotted SNP coverage depth and genotype concordance to determine the minimum SNP coverage required from the CNERs captures to achieve high concordance with the WGS genotypes. We find at least 10x read depth is needed from the CNERs enrichment data to obtain 95% concordance with the WGS data (Figure 3.5A). The read depth requirement varied based on the genotypes. To attain 95% concordance, homozygous reference sites need 12X coverage (Figure 3.5B) but homozygous alternative sites (Figure 3.5C) and heterozygous sites (Figure 3.5D) need at least 4X depth. It would be interesting to perform the complementary analysis to find the minimum depth required in the WGS data that produced high concordant genotypes with the CNERs data.

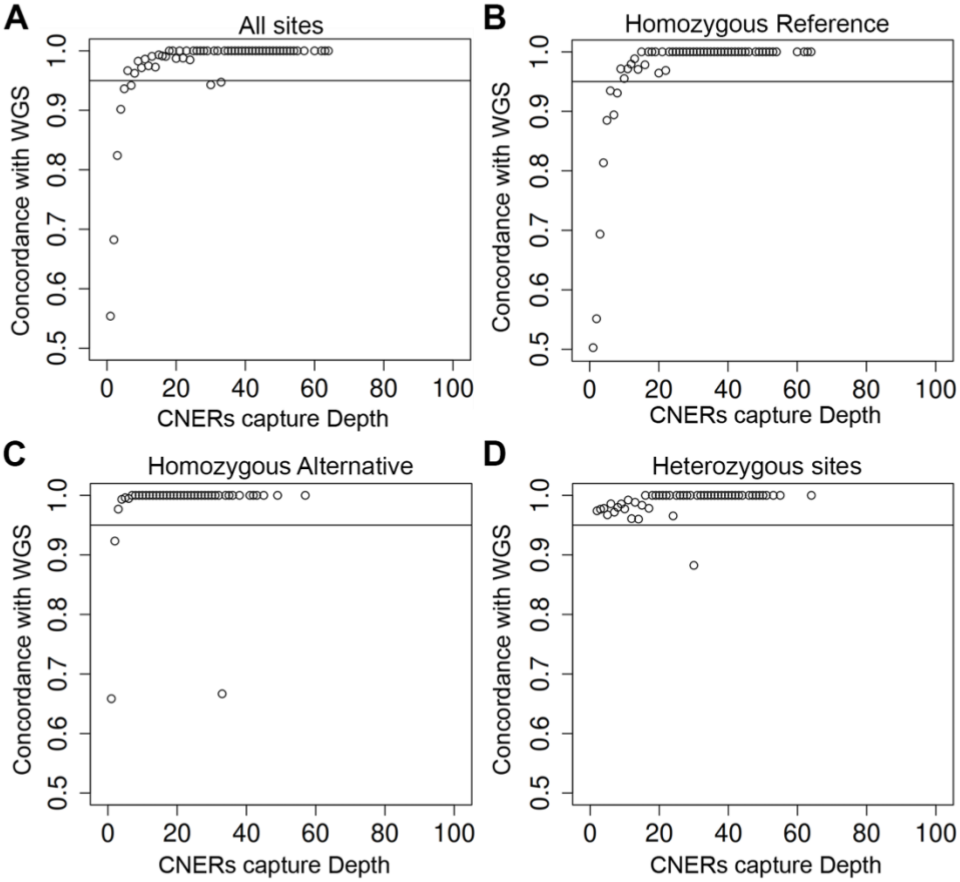


Figure 3.5 CNERs capture depth determine the genotype concordance.

*Genotype concordance between the CNERs enrichment data and the WGS data plotted against CNERs capture depth for all sites **(A)**, homozygous reference **(B)**, homozygous alternative **(C)** and heterozygous sites **(D)**. Vertical lines denote the 95% concordance.*

Discussion

We designed the CNERs panels that can be used for both degraded and fresh DNA samples and can be used for genotype profile generation at any stages of F/IGG. The Kintelligence assay is based on the multiplex PCR method [35] which is proven to be inefficient for degraded samples isolated from bones and hair specimens [29, 32, 34]. The panel was validated using DNA isolated from buccal swabs which is a preferred DNA source for profile confirmation stage of F/IGG, but not for profile generation needed for the initial search [2].

Previous SNPs panels designed for kinship analysis and DTC database search used SNP loci with more than two alleles having a MAF 0.1 - 0.9 [35–37], but we designed the panels to target biallelic SNPs with MAF 0.3 - 0.5. The FORCE panel was designed to target each autosomal SNPs with four RNA baits and two baits per X/Y SNPs. The 25K/95K panels were designed to target SNPs in genomic regions with <60% GC content due to the poor performance of RNA baits at higher GC regions [38, 39] (Chapter 1). The panel also made two or four RNA baits to target a SNP loci which increases both bait making cost and decreases total sequencing cost-savings.

We intentionally chose to target biallelic loci with MAF closer to 0.5, which are the hardest sites to genotype and require higher sequencing coverage to distinguish between homozygous and heterozygous sites (ref). We targeted each SNP with only one CNERs and demonstrated that our method efficiently captures target SNPs, produces higher coverage at the target SNP loci compared to the surrounding regions (Chapter 1).

With about 10M read pairs for a given hair sample, we achieved 46,455 (65%) target SNPs with at least one read coverage. Previous studies either covered a low percentage of target SNPs or used a higher raw reads to achieve high target coverage.The Kintelligence assay validated to perform with 60% call rate (~6000 target SNPs covered with one read) [35]. The FORCE panel recovered ~70% (~4,327) of targeted SNPs in the bone samples using ~41M reads [36]. The 25K panel recovered 54% (~13,406) targeted SNPs and the 95K panel recovered 39% (~36,972) from bone samples using ~5M reads [37].

Our results demonstrated that the CNERs SNP enrichment is cost-effective compared to the whole genome sequencing for genotyping forensic specimens. Further, compared to the previously designed SNP panels, we designed the CNERs panels to target biallelic SNPs with almost equal MAF which are the hardest sites to genotype. We also showed that CNERs would require a lesser number of raw reads to achieve a higher percent of target SNPs covered with at least one read. For these advantages, we expect the CNERs method, and the F72k SNP panel will be a valuable resource to the forensic community for F/IGG search and other forensic human identifications.

Declarations

Ethics approval and consent to participate.

This study was approved by the UCSC IRB Protocol #HS3382 to collect hair and saliva samples from the volunteers without any personally identifiable information. A copy of the Consent form to participate in the study can be found in the Supplementary text.

Consent for publication

The study volunteers consented to publish the results of this study when they consented to participate in this study as described in the UCSC IRB Protocol #HS3382. All authors read the manuscript, approved the results and conclusion and consented to publish this manuscript.

References

1. Greytak EM, Moore C, Armentrout SL. Genetic genealogy for cold case and active investigations. Forensic Sci Int. 2019;299:103–13.

2. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: Current methods, knowledge and practice. Forensic Sci Int Genet. 2021;52:102474.

3. Lu C, Greshake Tzovaras B, Gough J. A survey of direct-to-consumer genotype data, and quality control tool (GenomePrep) for research. Comput Struct Biotechnol J. 2021;19:3747–54.

4. Wickenheiser RA. Expanding DNA database effectiveness. Forensic Sci Int Synerg. 2022;4:100226.

5. Phillips C. The Golden State Killer investigation and the nascent field of forensic genealogy. Forensic Sci Int Genet. 2018;36:186–8.

6. Kennett D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. Forensic Sci Int. 2019;301:107–17.

7. van Oorschot RA, Ballantyne KN, Mitchell RJ. Forensic trace DNA: a review. Investig Genet. 2010;1:14.

8. Brandhagen MD, Loreille O, Irwin JA. Fragmented nuclear DNA is the predominant genetic material in human hair shafts. Genes (Basel). 2018;9.

9. Fischer H, Scherz J, Szabo S, Mildner M, Benarafa C, Torriglia A, et al. DNase 2 is the main DNA-degrading enzyme of the stratum corneum. PLoS ONE. 2011;6:e17581.

10. Dash HR, Das S. Microbial degradation of forensic samples of biological origin: potential threat to human DNA typing. Mol Biotechnol. 2018;60:141–53.

11. Alaeddini R, Walsh SJ, Abbas A. Forensic implications of genetic analyses from degraded DNA--a review. Forensic Sci Int Genet. 2010;4:148–57.

12. Hara M, Nakanishi H, Yoneyama K, Saito K, Takada A. Effects of storage conditions on forensic examinations of blood samples and bloodstains stored for 20 years. Leg Med (Tokyo). 2016;18:81–4.

13. Rahikainen A-L, Palo JU, de Leeuw W, Budowle B, Sajantila A. DNA quality and quantity from up to 16 years old post-mortem blood stored on FTA cards. Forensic Sci Int. 2016;261:148–53.

14. Fondevila M, Børsting C, Phillips C, de la Puente M, Consortium E-N, Carracedo A, et al. Forensic SNP genotyping with SNaPshot: Technical considerations for the development and optimization of multiplexed SNP assays. Forensic Sci Rev. 2017;29:57–76.

15. Voskoboinik L, Ayers SB, LeFebvre AK, Darvasi A. SNP-microarrays can accurately identify the presence of an individual in complex forensic DNA mixtures. Forensic Sci Int Genet. 2015;16:208–15.

16. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008;4:e1000167.

17. de Vries JH, Kling D, Vidaki A, Arp P, Kalamara V, Verbiest MMPJ, et al. Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. Forensic Sci Int Genet. 2022;56:102625.

18. Davawala A, Stock A, Spiden M, Daniel R, McBain J, Hartman D. Forensic genetic genealogy using microarrays for the identification of human remains: The need for good quality samples - A pilot study. Forensic Sci Int. 2022;334:111242.

19. Tillmar A, Fagerholm SA, Staaf J, Sjölund P, Ansell R. Getting the conclusive lead with investigative genetic genealogy - A successful case study of a 16 year old double murder in Sweden. Forensic Sci Int Genet. 2021;53:102525.

20. Tillmar A, Sjölund P, Lundqvist B, Klippmark T, Älgenäs C, Green H. Whole-genome sequencing of human remains to enable genealogy DNA database searches - A case report. Forensic Sci Int Genet. 2020;46:102233.

21. Kapp JD, Green RE, Shapiro B. A Fast and Efficient Single-stranded Genomic Library Preparation Method Optimized for Ancient DNA. J Hered. 2021;112:241–9.

22. Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet. 2015;18:78–89.

23. Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, et al. Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. Forensic Sci Int Genet. 2017;28:52–70.

24. Tillmar A, Grandell I, Montelius K. DNA identification of compromised samples with massive parallel sequencing. Forensic Sciences Research. 2018;:1–7.

25. Hollard C, Keyser C, Delabarde T, Gonzalez A, Vilela Lamego C, Zvénigorosky V, et al. Case report: on the use of the HID-Ion AmpliSeqTM Ancestry Panel in a real forensic case. Int J Legal Med. 2017;131:351–8.

26. Ambers AD, Churchill JD, King JL, Stoljarova M, Gill-King H, Assidi M, et al. More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing. BMC Genomics. 2016;17 Suppl 9:750.

27. Li R, Li H, Peng D, Hao B, Wang Z, Huang E, et al. Improved pairwise kinship analysis using massively parallel sequencing. Forensic Sci Int Genet. 2019;38:77–85.

28. Müller P, Sell C, Hadrys T, Hedman J, Bredemeyer S, Laurent F-X, et al. Inter-laboratory study on standardized MPS libraries: evaluation of performance,

concordance, and sensitivity using mixtures and degraded DNA. Int J Legal Med. 2020;134:185–98.

29. Sidstedt M, Steffen CR, Kiesler KM, Vallone PM, Rådström P, Hedman J. The impact of common PCR inhibitors on forensic MPS analysis. Forensic Sci Int Genet. 2019;40:182–91.

30. Zeng X, Elwick K, Mayes C, Takahashi M, King JL, Gangitano D, et al. Assessment of impact of DNA extraction methods on analysis of human remain samples on massively parallel sequencing success. Int J Legal Med. 2019;133:51–8.

31. Elwick K, Zeng X, King J, Budowle B, Hughes-Stamm S. Comparative tolerance of two massively parallel sequencing systems to common PCR inhibitors. Int J Legal Med. 2018;132:983–95.

32. Shih SY, Bose N, Gonçalves ABR, Erlich HA, Calloway CD. Applications of probe capture enrichment next generation sequencing for whole mitochondrial genome and 426 nuclear snps for forensically challenging samples. Genes (Basel). 2018;9.

33. Hwa H-L, Chung W-C, Chen P-L, Lin C-P, Li H-Y, Yin H-I, et al. A 1204-single nucleotide polymorphism and insertion-deletion polymorphism panel for massively parallel sequencing analysis of DNA mixtures. Forensic Sci Int Genet. 2018;32:94–101.

34. Bose N, Carlberg K, Sensabaugh G, Erlich H, Calloway C. Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples. Forensic Sci Int Genet. 2018;34:186–96.

35. Snedecor J, Fennell T, Stadick S, Homer N, Antunes J, Stephens K, et al. Fast and accurate kinship estimation using sparse SNPs in relatively large database searches. Forensic Sci Int Genet. 2022;61:102769.

36. Tillmar A, Sturk-Andreaggi K, Daniels-Higginbotham J, Thomas JT, Marshall C. The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications. Genes (Basel). 2021;12.

37. Gorden EM, Greytak EM, Sturk-Andreaggi K, Cady J, McMahon TP, Armentrout S, et al. Extended kinship analysis of historical remains using SNP capture. Forensic Sci Int Genet. 2022;57:102636.

38. Zhou J, Zhang M, Li X, Wang Z, Pan D, Shi Y. Performance comparison of four types of target enrichment baits for exome DNA sequencing. Hereditas. 2021;158:10.

39. Cruz-Dávalos DI, Llamas B, Gaunitz C, Fages A, Gamba C, Soubrier J, et al. Experimental conditions improving in-solution target enrichment for ancient DNA. Mol Ecol Resour. 2017;17:508–22.

40. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31:3555–7.

Appendix

Disclosures

Dr. Green and I are listed as co-inventors in a PCT application filed by the UCSC describing the methods presented in this thesis. I am a founder and shareholder of GenZ Genomics Private Limited in Chennai, India.