**Context:**

Horvath and Wilkinson profiled bats’ methylomes to discover age-correlated CpG sites and design DNA microarrays for age analyses. Our goal is to leverage their CpG site data to construct epigenetic age models specific to our three Madagascar bat species, and to develop a more sustainable, flexible methylation quantification pipeline than microarrays. We are in the process of designing a protocol—targeted bisulfite sequencing, potentially incorporating restriction enzyme genome reduction to further reduce costs.

**Identify targets:**

To build our species-specific epigenetic clocks, we need to first identify age-correlated CpG sites. CpG discovery is expensive—whole methylomes need to be sequenced at high depths to calculate methylation percentages and across many individuals to identify sites significantly correlated with age. We can skip whole methylome profiling by targeting the age-correlated CpG sites that Horvath and Wilkinson previously identified in phylogenetically related bat species. While we’re assuming that CpG sets are conserved across bat species, they won’t match up completely—we’ll need to translate their sites to our bat genomes. We want to find the subset of their age-correlated sites that can be identified in all three of our bat species so that we can have one protocol that targets the same set of sites in each species. The age models built from those sites will be species-specific, but by limiting ourselves to just one set of sites, we can avoid designing multiple sets of expensive probes.

To find this set of conserved age-correlated CpG sites, we should start by asking Horvath and Wilkinson for their **full set** of CpG-containing DNA fragments (i.e., age-correlated CpG sites flanked up- and downstream by ~150bp) that are conserved across bat species most closely related to ours: *Rousettus aegyptiacus*, *Pteropus* spp., and *Eidolon* spp. The aligner minimap2 (designed to find overlap between long noisy reads and a reference) can be used to identify which subset of their conserved sites map to all three of our bat species. We’ll then further filter this subset to sites flanked by enough conserved base pairs to be targeted by the same probe. This filtering process will give us a set of DNA fragments that are conserved across our three species, contain CpG sites that are likely to be age-correlated, and can be targeted by a single set of probes.

Lastly, we need to determine whether we can apply restriction enzyme genome reduction and still target a sufficient number of conserved age-correlated CpG sites. This approach would be a hybrid of reduced representation bisulfite sequencing (RRBS) and restriction-site-associated DNA sequencing with capture enrichment (RADcap). Restriction enzyme digestion saves time and funds by combining DNA fragmentation, end repair, and internal barcoding into a single reaction and increasing capture efficiency through pre-biasing the DNA pool for regions of interest before capture. We’ll run cuRRBS to in silico determine whether there exists an optimal combination of restriction enzymes that will recover at least ~2000 of our conserved CpG-containing probe-targetable fragments, as well as the size selection range we’ll need to recover that subset. Restriction sites must be appropriately distanced from CpG sites—close enough for fragments to fall within a reasonable size range, but far enough for to preserve the flanking conserved sequence for probe capture. Note that our final set of fragments may contain conserved CpG dinucleotides that are conserved across our three genomes but weren’t identified in Horvath/Wilkinson’s bats—our set of age-correlated sites will be some subset of the translated Horvath/Wilkinson’s sites + additional sites discovered through targeting their set.

Summary of target identification workflow: Horvath/Wilkinson’s age-correlated sites 🡪 conserved in species that are phylogenetically related to ours 🡪 also conserved in our three bat species and flanked by enough conserved sequence to be targeted by the same probe 🡪 able to be recovered via restriction enzyme cocktail = set of at least 2,000 targets (explain why 2,000 is next section—about probes).

Remaining questions:

* Does the genomic distribution of our target sites matter? Do we need to worry about many sites clustered on the same target fragment—because clustering increases the risk of functional links, which could skew our age model? Or is clustering to be expected because many sites will likely fall within CpG islands?
* I’m guessing Horvath and Wilkinson will have just one set of probes (i.e., one microarray) for multiple species...how did they go about identifying conserved regions? What was their cutoff for similarity?
  + Check out how other studies have used human methylation arrays on other species such as primates
    - Housman et al. (2019): aligned probes from the Infinium HumanMethylation BeadChips array to a primate genome--retained probes with e-value>10, 0 mismatches in the 5bp flanking the CpG site, and 0-2 mismatches in the 45 bp not including the CpG site. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6585300/>
    - Applied Human BeadChip array to primates: Mapped all 50bp probes, which were designed from the human reference genome, to three different primate species using BWA, allowing max edit distance of 3. Probe requirements:
      * Perfect match or 1-2 mismatches in the first 45bp and NO mismatches in the 3’ 5bp closest to the CpG site being assayed.
      * Human SNPs with minor allele frequency of <0.05 within the last 5bp of the probe binding site closest to the CpG.
      * For each species, SNPs with minor allele frequency <0.15 within the last 5bp of probe binding site closest to the CpG
      * <2 SNPs with minor allele frequency >=0.15 in the first 45 bp <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3764194/>

While we’re waiting for Horvath and Wilkinson to publish their CpG data, we can start by assessing CpG fragment conservation and potential for restriction enzyme recall across our three bat genomes…

**Identifying conserved CpG-containing fragments that can be recovered via restriction enzyme digestion in all three of our bat species:**

We began the Pteropus genome—extracting sequences for all CpG sites flanked up- and downstream by 80bp, and mapping these 160bp fragments to the Rousettus genome using the aligner minimap2. We selected all alignments with 60 or more contiguous matching base pairs (>= 60M in the CIGAR strings), thus removing short and unmapped reads (i.e., CpG-containing fragments that are not sufficiently conserved to share probes between genomes). We identified primary alignments for this set of conserved fragments using the minimap2 alignment (mapq) score, which rewards alignments that are long, contiguous, and unique or significantly stronger than alternative alignments for the given query fragment. Single alignments were assumed to be primary. Some single alignments still had low mapq scores—this is because many of our fragments are overlapping (for CpG sites located close together). Typically, we would remove overlap. However, we want to search the full set of possible CpG-containing fragments for those that can be recovered by restriction enzymes. So for now, we’ll keep low mapq scores as long as they’re unique or significantly better than the alternative alignments. We will remove any remaining overlapping fragments and re-check alignment scores once we’ve identified those that can be recovered with restriction enzyme digests and match the sites identified by Horvath/Wilkinson. With multiple alignments, we selected the alignment with the maximum mapq score when that maximum was high. However, low mapq scores across alignments suggested insufficient data to distinguish between potential alignments for the given query fragment. In particular, fully matching 160bp alignments shared a mapq score of 0 and could not be differentiated without more base pairs. In these cases, we extracted longer sequences for the corresponding CpG sites (flanked up- and downstream by 160bp) and reran the alignment—increasing flanking distance until we could confidently select primary alignments (by maximum mapq) for the majority of fragments. We mapped these alignments back to the Pteropus and Rousettus genomes, and ran cuRRBS to in silico determine whether there exists an optimal combination of restriction enzymes that will recover at least ~2000 of these conserved CpG-containing probe-targetable fragments, as well as the size selection range we’ll need to recover that subset. If we repeat these methods across all three of our bat species, we will have then identified a set of CpG-containing fragments that are conserved for single probe design and can be recovered via restriction enzyme digestion—we could theoretically just search for the set of Horvath/Wilkinson’s CpG sites that correspond to sites in these fragments.

**Probe design:**

Probes will be designed and manufactured with Arbor Biosciences. They strongly recommend post-bisulfite treatment capture, particularly you can’t amplify pre-treatment (PCR doesn’t replicate methylation, whereas bisulfite conversion encodes methylation in the DNA). Post-bisulfite capture requires 9 probes per target sequence—target variants for no sites methylated, all sites, 50% of sites, and other 50% of sites on both strands, as well as the original untreated sequence). 20K probes only targets 100kb of sequence, or 1000 (if 100bp probes) to 2000 (50 bp probes) sites. EGL multiplexes 48-96 samples per capture—the 48 reaction kit would be sufficient for 2304 - 4608 samples. We may be better off with the 16 reaction kit because we currently have only 1,291 wing punch samples. How will we decide probe size?

**Library preparation**

For now, let’s assume we’re going to apply an RRBS approach:

1. Enzyme digestion + ligation of stub adapters with barcodes (borrowing the RADcap approach (1): Restriction enzymes and short forward/reverse adapters with internal sample barcodes are added in the same reaction, and the samples are incubated. Adapters must be **methylated** and specific to each restriction cut site (protocols for producing methylated adapters?). After incubation, ligation mix is added and temperatures cycled to promote adapter ligation and subsequent digestion of chimeras and dimers.
2. Pool
3. Bead-based size selection: optimal size ranges determined in silico using cuRRBS.
4. Bisulfite conversion
5. Indexing PCR: iTru5-8N sequence tags (indices) for de-duplication, gets double-stranded DNA before capture
6. Capture
7. Single-end vs. paired-end sequencing: relatively common to use single-end sequencing because read overlap can skew methylation percentages (2,3). Though there must be bioinformatics pipelines to combat this given we have a reference genome? Dark sequencing to compensate for lack of variation in beginning bases of MspI-digested fragments (first three bases are non-random: CGG) (4)

If we can’t use RRBS, Kapa is better than Nextera (despite separate fragmentation step). Unless we can homebrew Nextera because tagmentation is preferable. Looking into growing our own transposase…

Random notes for me:

* Enrichment methods: balancing sensitivity (% of targets captured), specificity (% on-target), and coverage uniformity—how does pre-biasing the pool help this??
* When calculating how many reads you need to get the depth you need per sample, you have to consider the percent off-target reads and percent duplication rate (happens during PCR—sounds like it can be caused by too much input material?)
  + And maybe even bisulfite conversion rate? Though that seems to be >99%
* Calculating coverage requirements: Ziller et al. 2016. Targeted bisulfite sequencing of the dynamic DNA methylome explains how they did it
* Unmethylated control sequence—to determine the efficiency of the sodium bisulfite conversion reaction
* Long adapters can interfere with each other during enrichment
* Clarify our own definitions of read depth vs. coverage
* Read depth: 200X?
* Bisulfite treatment significantly reduces sequence complexity AND degrades DNA
* Calculate bisulfite conversion efficiency based on unmethylated cytosines in the end repair position: identify sequences that contain the adapter on their 3’ end and count up the number of times the filled position was not converted to T = percentage of non-bisulfite converted cytosines in the fill-in position (from end repair step) (5)
* Check for adapter contamination—especially if have long reads (stringent adapter removal in bioinformatics pipeline—even a single bp overlap with a adapter sequence)
  + Adapter contamination is particularly detrimental to methylation analyses (5)