Here’s the structure of the P1 and P2 adaptors, taken from Peterson et al 2012

P1 flex adaptor, contains sample barcode (NlaIII overhang)

For the flex adaptors, the overhang is on the P1.1 oligo, not the

P1.2 oligo seen in the Peterson sup mat.

P1.1 5’ ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcatgCATG 3’

||||||||||||||||||||||||||||||||||||||

P1.2 3’ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAcgtac-PHOS 5’

P2 adaptor, no barcodes (MseI overhang)

P2.1 5’ GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT 3’

|||||||||||||

P2.2 3’ BIO/AACAAGAGCGAGAAGGCTAGATTAA/5Phos 5’

Green sequences are the recognition sites for the PCR amplification step

What this looks like for a full RAD tag containing genomic DNA. The read 1 and read 2 sequencing primer sites are underlined.

5’ ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcatgCATGNNNNAATTAGATCGGAAGAGCGAGAACAA/BIO 3’

|||P1 ADAPTER |||||||||||||||||||||||| ||P2 ADAPTER

3’ TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAcgtacGTACNNNNTTAATCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5’

In deciding where to insert the degenerate base regions (DBR) we need to (1) not mess with the overhangs (in red), (2) not mess with the barcodes (in blue), (3) not mess with the PCR start sites (in green), and (4) not mess with the sequencing primer sites (underlined). That basically leaves 3 spots:

1. In the P1 adaptor, “left” of the barcode
2. In the P1 adaptor, “right” of the barcode (between the barcode and the overhang).
3. In the P2 adaptor, “left” of the read 2 sequencing site (between the sequencing site and the overhang.

5’ACACTCTTTCCCTACACGACGCTCTTCCGATCT**1111**gcatg**2222**CATGNNNNAATT**3333**AGATCGGAAGAGCGAGAACAA/BIO 3’

|||P1 ADAPTER ||||||||||||||||||| ||P2 ADAPTER

3’TGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA**1111**cgtac**2222**GTACNNNNTTAA**3333**TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG 5’

To my mind, the main choice is between putting it in the P1 vs the P2.

P2 makes more sense to me: we only have to buy one new set of primers (instead of 48). As long as we do paired-end sequencing, we’ll be able to detect PCR duplicates. As far as I understand, the algorithm for detecting PCR duplicates is basically:

1. demultiplex the reads from the barcodes.
2. Within each sample (within each barcode), find reads that are identical. For PE reads, this means finding sets of reads that are identical for R1 and R2
3. For each set of identical reads, look in the DBR region. Reads with different DBR come from different templates. Reads with the same DBR are PCR duplicates, and all but the highest-quality read should be discarded.

This sort of thing can be implemented in Python or Perl scripts (the published ones I’ve seen are for Python, I believe), and my guess is that it would be fairly simple to modify such scripts so that they look for the DBR in the P2 adaptor.

I can’t think of any strong reason why have the DBR in the P2 adaptor should make a difference, as long as we do PE sequencing (with single-end sequencing we’d need to have the DBR in the P1). It may mean we can’t check for PCR duplicates when the R1 and R2 reads don’t match up, but that happens pretty rarely, and may be a sign that we want to discard those reads anyway.

This implementation that I’ve described above is slightly different from the one in Tin et al. 2015: they do a double digest, but not the “true” ddRAD protocol from Peterson. I wrote up all this, and then Marco showed me this neat paper (Schweyen et al. 2014) on using DBR regions to avoid PCR duplicates in ddRAD. It’s hidden away in a small journal (The Biological Bulletin) and has only been cited once (and not in empirical usage). But, they detect PCR duplicates using DBR in the true ddRAD framework, and they put their DBR in the P2 adaptor (location 3). Their DBR design is a bit more complicated than in Tin et al.: instead of 4 random bases, they add 10 base pairs to the adaptor. On the strand with the overhang, they add (5’ to 3’) 2 Cs, 3 “I”s, and 5 Ns. “I” stands for 2’-deoxyinosine, a base that binds with all 4 bases, preferring in decreasing order C, A, and T. On the other strand, they have two Gs, two M, one H, and 5 Ns (where M is A or C and H is A, C, or T). This design is, according to them, meant to reduce mispairing of the top and bottom strands during annealing. I sort of get why it would do that, but I wonder why you wouldn’t just use I for all of the bases in the DBR, not just 3 of them. They also use a slightly different annealing procedure than Peterson et al. But, they incorporate their library index into the P2 as well, instead of adding it during PCR, so that might be why. We could maybe get advice from IDT about the best practice for annealing.

In our scheme, the P2 adaptor would look like this on the fragment:

5’ GENOMIC\_NNAATT**CCIIINNNNN**AGATCGGAAGAGCGAGAACAA/BIO 3’ Strand with overhang on adaptor

3’ GENOMIC\_NNTTAA**GGMMHNNNNN**TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG Strand with PCR site and read site

Interestingly, their primer design is actually a little more complicated than that: on both their P1 and P2 they add a little insert of 1-3 bases between the PCR overhang and the start of the DBR to prevent “blinding” of the sequencing lasers when they read the cut sites, which are the same across sequence fragments. No idea if that’s something we should be worried about, I’ve never heard it mentioned before in any publication on RAD (or . We could possibly implement something similar by ordering 4 different P2s with either 0-3 bases in the insert, such that ¼ of reads are offset by a base and maybe there is less “blinding,” but not sure if that is worth it.

In summary, I think we should asically copy the Schweyen et al. DBR design, by putting the DBR in the P2 adaptor and using their scheme of degenerate bases and alternative nucleotides, but not bother with the inert thing. Bioinformatically, there are a few options we can use to remove PCR duplicates, but all of them would likely require some customization.

Would be great to get some input from you all. Would also be good to run some of this by people with a bit more ddRAD/adaptor design knowledge. Specifically:

1. Does this scheme make sense?
2. I want to confirm that the strand with the “I”s should be opposite of the PCR start site. Because of the forked adaptor design, initial PCR should only start from the P1, so the I’s won’t be read in the initial PCR site. Marco and I went over this and he was a bit unconvinced, so I want to make sure I put it on the right side (I did the same thing as Schweyen et al).
3. Why not use the “universal base” I of all the DBR sites in the non-PCR P2 strand? Just because it binds preferentially to C, A, and T? I’m getting some advice from IDT, our oligo supplier, about this, annealing temps, and the best way to order the primers.
4. Has anyone else heard of “blinding” the laser on an Illumina sequencer (I’m emailing genome Quebec to ask as well)?