# GTseq locus selection and primer design

Written by Karen Martien, based heavily on info provided by Anthony Clemento

[July 20, 2023: This is a work in progress. I haven’t bothered to fill in any details in the SNP Discovery section. Those methods are pretty standard, so most people are probably either familiar with them or can ask me for help. The Locus Selection and Primer Design section is more-or-less complete, though I’ll probably think of other things.

I’ve just finished (hopefully) the panel optimization for humpback whales, and am awaiting production data. I may update the Panel Optimization section once I have those data. I also intend to add sections detailing how to go from fastq files coming off the sequencer to QA/QC’ed genotypes ready to analyze, but that will come after I’ve actually completed those steps. Getting from a fastq to called genotypes is particularly tricky during panel optimization, because you don’t yet \*really\* know which loci/samples are working, so it’s hard to set thresholds for, e.g., eliminate loci that produced genotypes at <X% of samples or samples that are genotypable at <Y% of loci. So, I’m still figuring out what will be the most useful pipeline for that processing, and will update all aspects of this document as necessary once I figure that out (though I’m pretty sure that the fastq file full of Ns noted in the first paragraph about Primer3 is going to be clutch in that process!).]

Required starting data:

1. A reference genome
2. Some type of RAD data (ddRAD/RADseq/2bRAD)

In some cases, I have R scripts that accomplish various steps. Those scripts are noted in *italics* in the relevant places. For the most part those scripts are not written in a way that makes them generally applicable to other projects, but they at least provide considerably more step-by-step detail than I’ve covered here, and at best some may need only minor modification to make them useful to others.

## SNP Discovery

1. Trim
2. Align
3. Variant calling
4. SNP filtering

## Locus Selection and Primer Design

1. Identify potentially informative loci

Select SNPs that are likely to be the most informative for your question. You’ll eliminate a lot of SNPs in subsequent steps, so select ~4x as many SNPs as you’d like in your final panel. Metrics to consider when evaluating SNPs:

* Overall heterozygosity
* Heterozygosity in a focal stratum
* Importance of each SNP in a Random Forest analysis used to assign samples to strata
* Groups of SNPs sufficiently close to each other than they can genotyped in a single locus

1. Design primers for selected SNPs (*Primer.design.R*)

Design primers for use with 150bp Illumina paired-end sequencing. A schematic for a target amplicon looks like this (example is max primer length and max target length):

|<-- upstream -->|<-- 21bp primer region -->|<-- target max 105 bp -->|<-- 21bp primer region -->|<-- downstream -->|

This leaves an overlap of 3bp between fwd (75bp) and rev (75bp) reads which is used to flash the reads together into a single fragment - more overlap is desirable (which happens when primers are shorter than 21bp and the target is shorter than 105, in this example) but this is the max. Anthony Clemento et al. experimented with single-end sequencing (150bp) but the quality was bad compared to paired-end, as the flash step is a psuedo-filter for quality (low quality reads will not flash together).

We are not interested in the variation upstream or downstream of the primers, and of course we desire no variation in the primer sequences. Ultimately, we are looking for ~100bp windows that maximize the number of SNPs, but still allow for primer design.

For primer design we use Primer3 (<http://bioinfo.ut.ee/primer3/>) or BatchPrimer3 (<http://batchprimer3.bioinformatics.ucdavis.edu/cgi-bin/batchprimer3/batchprimer3.cgi>). The input required by Primer3 is a fasta file with all of your potential target amplicons with all variable sites converted to Ns (*Primer.design.R* generates this file). The SNP for which you want to design primers should be bracketed with [...]. You can manually edit the fasta file so that the brackets enclose multiple SNPs if you want to force Primer3 to look for primers that amplify multiple SNPs. SAVE THIS FILE! You’ll need it during data analysis (though you’ll need to delete the brackets).

Leave the default Primer3 settings, with the exception of changing the Product Size Ranges to 90-143. If you notice that you're not getting enough to successfully design, you can stretch both the Primer Tm range (min:56.0, max: 63.0) and the Primer GC% (min:25.0, max: 75.0). You can also investigate shorter primer lengths, but not less than 16 bps.

If Primer3 identifies multiple primer pairs for a target, it will return them in rank order, with the first one being the pair it considers the ‘best.’ In terms of selecting the ‘best’ primer pair from Primer3, try to get as many as possible to have products of about the same size (narrowing that 90-143 range, but this has proven to not be that big of a deal) and then avoid primer pairs with warnings of potential secondary structure (also not that big of a deal as what Primer3 generally spits out tend to work well). The strategy I used for humpback GTseq was to choose the highest-ranked primer pair that had a product less than 130 bps long, unless that pair had more warnings than a pair with a longer product.

1. Evaluate primers/amplicons for potential problems

* For loci on the same chromosome, see how far apart they are. Ideally they should be at least 100,000 bp apart (*Primer.design.R*).
* Use BLAT or BLAST to compare all of your target amplicons to each other. Amplicons with a high degree of similarity may interfere with each on the panel, so avoid them (*blast primers.R*).
* Map your ddRAD/RADseq/2bRAD data back to the amplicons you’ve designed. Do they map well? Run the alignments through FreeBayes - are all of your target SNPs actually identified as high-quality SNPs? Do some loci have dramatically more mapped reads than others? Maybe those are going to high rollers in your GTseq data set too.
* Map your primers to your genome. Look for primers or primer pairs that map to more than one location. (*Evaluate.primer.alignment.R*)
* Map your amplicons to your genome. Look for loci that map to more than one location.

## Panel Optimization

Eliminate loci that:

1) get no/very few reads

2) get orders of magnitude more reads than the other loci (i.e., the ‘high rollers’)

3) have a large imbalance of forward to reverse reads. Check this by counting the number of times the forward and reverse primers for each locus appear in your fastq files (*0\_Start.here\_Summarize.fastq.files.R*). If you eliminated loci that had a primer that mapped to your genome in multiple places (bullet point 4 under Locus Selection – 3. Evaluate primers/amplicons for potential problems), then all loci should have roughly equal numbers of forward and reverse reads. Any locus where one primer has >20x as many reads as the other should probably be eliminated.

Anthony Clemento recommends starting with panels of 96 loci, kicking out problematic loci, then repeating the experiment. Once you get down to 60 to 70 remaining loci, combine them into panels of 120-140 loci and again check for problem loci. Performance of some loci can change dramatically when other loci are added or removed from the panel.

For humpbacks, we started with a single panel of 528 loci. GTSEEK eliminated 144 based on criteria 1 and 2 above. I had NOT mapped my primers or amplicons to my genome during locus selection, so I had 16 remaining loci with large imbalances in their forward and reverse reads (#3 from above). [Note that I then, belatedly, mapped my primers and amplicons to the genome and found that a large fraction of the loci eliminated from my panel during optimization would have been eliminated during locus selection if I’d done the mapping then.] We’re moving forward with the remaining 368 loci, even though some of them are not producing enough reads to be usable.