Sorted Barcode Sequencing (SoBaSeq), or low-input dimer-less amplicon sequencing of strain barcodes in deletion collection

The purpose of this protocol is to allow you to get "barseq" libraries (amplicon sequencing of barcodes in the yeast deletion collection) out of very low input genomes. It seems to work, when applied to a FISH FACS Seq (FFS) application (each reaction was 1/3 of gDNA extracted from ~5e5-1e6 PFA fixed cells via overnight reverse-crosslinking and phenol chloroform extraction).

Shopping list:

- Enzymes and other reagents
  - Vent (exo-) DNA polymerase (NEB M0257) & included 10x ThermoPol buffer and 100mM MgSO4
  - NEB dNTP mix (ND447S diluted with hyclone H_20 to 2.5mM)
  - BSA (NEB 20mg/ml B9000S)
  - 50% glycerol, diluted from 100% (Fisher) with hyclone water
  - exoI from Thermo (EN0581)

- Primers, 5' to 3', in IDT syntax
  - DGO1562: GTCTGAACTCCAGTCACATCNCNCNCNTNCNGTCGACCTGCAGCGTA
  - DGO1588: CCATTGGTGAGCAGCGAAGGATTTGGTGGA/3Phos/
  - DGO1589: agaaaaagcagcgtaGATGTAGAAGCAAGA/3Phos/
  - DGO1567: GATGTCCACGAGGTCTCT
  - DGO1576: CGTACGCTGCAGGACGCAC
  - DGO1519: CAAGCAGAAGACGGCATACGAGATGTCTGAACTCCAGTCAC
  - Forward index primer for index barseq (on a plate), ours looks like: ACCGCTTCTCCGATCTNNNNNGTCCACTCCAGTC
  - Ours are a set of 120 different 5bp combinations. If you're making new ones, be mindful of spacing them further (in terms of mismatches away from each other) and use variable length to help space out the fixed sequence to take it easy on the Illumina base calling software.

- Tubes and gear
  - Lots of single PCR tubes, some tube strips for additivies
  - Multichannel that can do 1.2ul (the 2-20 does it okay)
  - Tiny centrifuge for PCR tubes, PCR tube sized rack, etc.
  - Agarose gel box and associated supplies for pouring gels, 100bp ladder

Some notes:

- Gels are all 3% agarose wide comb gels, ran at 100V for ~25min, poured with SybrSafe dye in-gel and imaged for long exposures on a nice gel box.
- To make a 3% gel, add ingredients to flask, then add ~20% more water (2.5% gel), then boil off the excess water (by weight).
- The choice of thermocycler matters! This was successfully run on BioRad T100's. The specs (http://www.bio-rad.com/en-us/product/t100-thermal-cycler) say it'll ramp at 2.5 average and 4C/s max.
- The choice of tube matters! This was optimized using single PCR tubes (round top). We found that it didn't work in a 96 well plate (probably bad sealing), so we just ran our 90 samples in three batches of 30 single tubes. Not that difficult.
- For keeping single tubes organized, I recommend a multi color sharpie striping scheme.

Points of interest:

- This incorporates a UMI during the first round of extension. This is an interspersed UMI, so it alternates fixed and degenerate bases to reduce dimer formation.
- This uses an exol step at 37C to digest the UMI, and the extra reverse primer that generates a lot of dimers.
- To block non-specific genomic targets and also dimers, 3' phosphorylated blockers are used to out-compete these.
- Glycerol is added to help the forward primer un-hairpin and actually prime efficiently.
- Note that Vent (exo-) is reported to maintain reasonable fidelity at higher Mg concentrations, so be cautious about lowering this.
- This uses 40x of amplification in the middle of the protocol, but note that the forward primer is limiting at 1nM. So this should plateau out there, and should be a way to balance concentration across multiple libraries. Seems to work.
The p5 addition is a separate reaction because the index primer will out-compete it's hybridization to the landing pad provided by the index primers. So keep it separate to avoid more dimers.

Future possible optimization directions, for the ambitiously selfless scientist:

- This is a total of 52 PCR cycles in the first set of reactions. There aren't all "real" generative cycles (limited by 1 nM primer to normalize across libraries), but they might introduce mutations or chimeras. This could be toned down, so long as there's a good way of measuring successfully amplicon generation ( tapestation perhaps?). If so, you'd have to be more careful with pooling at the end, to make sure you don't dilute out some libraries too much.
- With fewer cycles, you should try longer denaturation times to see if you reduce GC bias in the PCR.
- UMI specificity - if you omit the 1st extension reaction, you ought to not get products, but you don't. Presumably this comes from UMI primers sticking to other places in the genome, or otherwise just getting protected from the exoI digestion, and it seems that we get about 1/20th the product when we omit the 1st cycle. So a different UMI does not mean it's a different template, but it's a strong suggestion. There might be a way to get this to work better, but you'd need to balance the heat to denature these off-targets against denaturing the exoI activity. Maybe adding the 3' phosphorylated blocker oligo before denaturing the exoI would allow you to block re-priming?
- Are all the additives and steps necessary? Each one was added because it made a small incremental improvement at the time of addition, but the extent of epistasis has not been measured.

I recommend this series of work to build up to successful application of technique:

1. Get some gDNA from a BarSeq experiment you know worked, so gDNA of the KO collection pool (or singles if you want to verify this more?). Quantify with qubit to estimate about how many genomes you have. Dilute to make a standard.
2. Try the round1 PCR with "DGO1562" and "DGO1567" at 200nM concentration using a range of gDNA inputs (1e3, 1e4, 1e5, 1e6, 1e7 if you enough). You should see a product forming at 1e6 (around 100bp), and lower concentrations will make dimers (smaller). Be mindful that these BarSeq libraries run slower than their theoretical, presumably because they're complex libraries and the centers aren't going to hybridize as dsDNA.
3. Try doing the full procedure, but only one index primer for ease. Use a dilution series of gDNA of 1e1, 1e2, 1e3, 1e4, 1e5, or water. Try doing the round1 melt, anneal, extend (positive) or leaving it on the bench (negative). Compare the formation of products with or without the initial extension. You should see the negative shift to lower sensitivity, but that it is not perfect.
4. Then, try the full procedure on a few mock samples with a few indices, using the same pipettes and timing as you will need for the actual procedure. Try allowing yourself 3-4min for repriming. This should be as much of a dress rehearsal as is practical, and you should do this to make sure you're not going to waste your samples.
5. Then, complete the full technique.

The method:

- Extract gDNA however you like, and re-suspend in hyclone H_2O.
  - I recommend you do this in ~35ul so you can do a triplicate of each PCR reaction (keep in mind counting UMIs is very much not the same as counting absolute # of amplicons, as UMIs are not 100% sensitive or correct ). You're going to want as much as possible, but this has been done successfully (we saw bands) on samples with ~1e6 cells sorted into tubes, with a ~30-40% efficient extraction by weight, phenol chloroform step, then split each way, so an optimistic 1e5 genomes but pessimistic 1e4 genomes input. I've not tested inputs above that, but I assume it'll work. I've even got bands out of ~300 genomes (estimated by qubit and dilution), but I would not recommend trying that unless you're sure you just want presence/absence metrics.

- Make a dilution of your indexed forward primers to 10uM. I mention this because it's a bit time consuming, I assume all other reagents are diluted appropriately.

- Prepare master mixes when appropriate (R1, exoM, R2 now; R3 later)
  - 1x "R1" round 1 master mix ( 10ul master mix per sample, 20ul reaction volume, the below recipe is to make a doubly-concentrated mix so you can add 10ul gDNA and it's right concentration in your PCR tube)
    - 4.2ul hyclone H_2O
    - 2ul 2.5mM each dNTPs
    - 2ul 10x ThermoPol buffer
    - 0.6ul 100mM MgSO4
    - 0.2ul 20mg/ml BSA
    - 0.2ul 10uM DGO1562 "R1R"
    - 0.2ul 1uM DGO1588 "R1B1"
    - 0.2ul 1uM DGO1589 "R1B2"
    - 0.4ul Vent (exo-) polymerase
  - 1x "exoM" exonuclease I master mix
    - 0.25ul exoI (Thermo 20U/ul)
    - 1.75ul hyclone H_2O
  - 1x "R2" master mix ( second round of UMI'd barcode amplification, 5ul total per sample )
    - 0.51ul hyclone H_2O
    - 0.7ul 10x ThermoPol
    - 2.7ul 50% glycerol
    - 0.21 100mM MgSO4
• 0.07 20mg/ml BSA
• 0.27ul 10uM DGO1576 "R2B"
• 0.27ul 1uM DGO1567 "R2F"
• 0.27ul 10uM DGO1519 "R2/3R"

• 1x "R3" master mix ( actually make this later! during R2 cycling )
  ( this is for adding on the sample index primer, make a master mix without the index primer, then mix 4.8ul of the master mix with 1.2ul of the index to make twice what you need in strips of PCR tubes)
  • 0.2ul 10x ThermoPol
  • 0.6ul 10uM DGO1519 "R2/3R"
  • 0.6ul 10uM indexed forward sample primer, different one for each sample
  • 1.6ul hyclone H_2

• Put enough of the R2 and exoM mastermixes into PCR tube-strips, put on ice.
• In batches of 5 tubes, put 10ul of 1x R1 master mix then 10ul gDNA into the tube, then put it on ice.
• When all are done, spin it down and start the run on a T100 thermocycler, as set for a 30ul volume with lid heat set for 95C
  • 95C 4min to denature genomic
  • 50C 30s to anneal the R1R
  • ramp to 72C at 1C/s, hold for 30s
  • 37C hold
• With haste, open the thermocycler when it hits 37C. Open tubes in sets of 5, add 2ul exoM mastermix to samples and mix ~10 strokes with a multichannel. This contains glycerol, so it does not mix so well. Close tubes, continue. When all done, close lid and advance program.
  • Incubate 37C for 20 minutes to digest the R1R primer.
    @~10min into incubation open lid, then two tubes at a time take out, do a little (little) vortex to get the reaction all over the walls, then quick spin it back down, put it back on cycler.
    When all are done, then close lid and complete incubation.
  • 50C 5min to denature things a little and start to kill the exol
  • 80C 5min to kill the exol
  • 60C hold
• With haste, open the thermocycler. In sets of 5, add 3ul R3 and mix 2 strokes with pipette. Close and continue with program.
  • Prepare the 1x "R3" master mix now! ( Since you have a lot of downtime there. ) Fresher mixes! And it focuses your thoughts.
    • 95C 1min denature
    • 40 (!!!) cycles of
      • 95C 15s
      • 54C 15s hybridizing for 1567 forward, and 1519 reverse
      • 72C 20s
    • Hold 60C
• With haste, open the thermocycler. In sets of 5, add 3ul R3 and mix 2 strokes with pipette. Close and continue with program.
  • 95C 15s denature
  • 50C 15s should be index primer sticking onto the abundant minus strand products from 1519 extension
  • 72C 15s
  • 12 cycles of
    • 95C 15s
    • 68C 30s index and 1519 hybridization
    • 72C 30s
  • 4C hold, but of course you can take them to ice before it hits 4C
• Put samples on ice.
• You should have ~26ul of reactions left. You can run 10ul of this on a gel (3%) to check sizes. If you've done the preliminary experiment controls before, you'll know how big the products should look, and what a dimer looks like:

Left is 100bp NEB lader, so bottom rung is 100bp, then 200bp. Smear at top of lanes is the BSA. Note the third lane on the top is a little light on the reaction, and we saw this oddness in the PCA after sequencing for all PCR replicates of this sample, so I think that's a bad
gDNA extraction. Anyways, this should be about a 128bp product (+/-), but it runs slower than that.

You can use this gel to estimate how concentrated your libraries are, roughly. I recommend you classify them into a small number of pools based on observed or expected input amounts (like "input", "high gates", "low gates") so you can normalize these separately for the sequencing run, just in case.

- You can freeze your samples, or continue.
- When all your reactions are done, then pool (as described above) into a small number of pools. Keep this cold. Clean this up on a minelute column, and elute in appropriate volume (20-30ul?).

- Quantify this pool. As a control, submit some for Sanger sequencing using the DGO276 (Illumina P5 adapter) above as a primer. This works well for Genewiz. Here's what it should look like:

Note the 5bp degenerate at the start, then we see the fixed sequence emerge. After the TCTCT (red blue), we see the strain barcodes. This is abiview, so red is T, blue C, green A, black G. Notice the GC bias in the strain barcodes, likely the result of incomplete denaturation of PCR. The A spike in the middle is normal. Then it's fixed sequence (but some amplicons are not 20bp, thus out of phase), then you see the interspersed UMI (C,G,A,G,G,G interspersed with degenerate). Great.

- Prepare a master mix for each reaction for adding the p5 adapter:
  - 1x master mix "R4"
    - 9.95ul purified template + hyclone water
    - 1.2ul NEB dNTPs 2.5mM each
    - 1.5ul 10x thermopol
    - 0.45ul MgSO_4
    - 0.75ul 10uM DGO1519
    - 0.75ul 10uM DGO276
    - 0.4ul Vent (exo-) polymerase
  - (How much template to add? I've done ~60nM final and had good results, but you might want to add more? You will probably have enough template to try this part a couple of times. Maybe try a 200nM template final concentration?)

- Run this mixture on the PCR machine. For this step, I used an older thermocycler, but I assume you can use a T100 again. I would recommend limiting the ramp speed to 1C/s, just in case.
  - 95C 1min
  - 52C 30s
  - 72C 1min
  - 95C 30s
  - 68C 30s
  - 72C 1min
  - 4C and ice

- Run this on a 3% agarose gel. I ran it all and gel purified the bands later, but I ended up cutting out all the bands so I don't think this is a helpful step. You can probably just column purify it for sequencing. Here's an example of a gel:
The ladders are NEB 100bp ladders. The middle four are four different pools reactions. The far right band is the far right of the four pools, but the reaction is missing DGO276 (the p5 adapter). We reason that after the reaction there should be a mixture of products:

- **Strands:**
  - 1x 128bp + strand (original template)
  - 3x 128bp - strand (original template + two products from DGO1519 priming off template)
  - ∼2x 171bp + strand (< two products from DGO276 priming off template)
  - 1x 171bp - strand (from DGO1519 priming off of the DGO276 product)

- **Configurations:**
  - 128bp dsDNA
  - 128bp ssDNA, probably runs higher than dsDNA (far right lane upper band)
  - 171bp dsDNA
  - 171bp ssDNA, probably runs higher than dsDNA upper bands
  - 128bp/171bp ~dsDNA hybrid, ditto

- Anyways, then you can clean these up separately, then roughly estimate the concentration using qubit. Keep in mind it's not dsDNA, mostly.

- Do qPCR quantification of your pools, and sequence them (1x75bp Illumina)!

qPCR (Kappa kit) seems to be an accurate measure of the products that will actually stick to your flow cell, as assessed by flow cell density & GenCore Wisdom.

Beware NextSeq! It has a hard time discriminating amplicons from each other when seeded at the normal concentration. We had a 400e6 read run seeded at 5% phiX that only returned ∼30% of expected reads. GenCore said the micographs looked well clustered, but base-calling got very confused during the fixed sequence sections. Of the ones that worked, cycle 10 was a random base generator. A later run on NextSeq with conventional Barseq spiked with 15% phiX still had quality problems, but much better return of data. Recommend ∼25% phiX and low seeding to start conservatively, then get more aggressive from there.

- For analysis, see the forthcoming analysis pipeline [SLAPCHOP (with UMAMI), pheniqs or starcode+perl script, bwa mem, umi_tools]