

## Frequently Asked Questions

### 1. How will my sample be sequenced, and how much material do I need?

**Genomes.** Samples submitted for genome sequencing using 454 Titanium pyrosequencing technology will target a sequence coverage of 30X for genomes <250KB in size (Class-I) and 15X for genomes >250KB in size (Class-II). The Broad will accept Class-I samples prepared without the use of a purification step (e.g., CsCl, sucrose gradient, nuclease treatment, etc.); Class-II samples will require purification consistent with the research objectives. In addition, the absence of contaminating bacterial DNA must be confirmed (and documented upon sample submission with a gel image) in Class-II samples using PCR targeted at host markers such as the 16S/18S rRNA or recA genes.

At minimum, 100ng of template DNA should be submitted. If 100ng of DNA template is not obtainable from your sample, lesser amounts will be accepted pending consultation with the Broad. We anticipate that the target coverages indicated will capture >99% of the genome and generate a single contig assembly of your strain. Given the unknown nature of many genomes submitted as part of this project the following considerations may impact final assembly and result in more than one contig and/or a lower percentage of the genome captured: (i) levels of host contamination in excess of 50%, (ii) the presence of multiple viral genomes in a single sample, and (iii) inaccurate reported genome size estimates, and (iv) extremely low template quantities. The Broad will review such cases with the GMBF and determine if required resources are available to improve the genome assembly.

All samples need to be submitted as ds-DNA templates. The Broad will accept ds-DNA and ds-cDNA templates generated from ss-DNA and RNA viruses respectively. The generation of such DNA and cDNA templates is the sole responsibility of your research group. Such DNA and cDNA templates will require fragment sizes of either >2kb or 500-700nt, respectively, to be considered for genome sequencing. While samples from RNA and ss-DNA viruses that meet these specifications will be sequenced, genome assembly and annotation results for these samples are not certain.

**Viromes.** Virome samples will be sequenced using ¼ of a 454 Titanium pico-titre plate delivering an average of 200K reads with an average read length of 400nt. At minimum, 100ng of template DNA should be submitted. Virome samples are required to undergo a purification step consistent with the research objectives (e.g. CsCl, sucrose gradient, nuclease treatment, etc.) prior to submission. It is required that researchers confirm the removal of bacterial/other contamination by PCR targeting markers such as the 16S/18S or recA genes. All samples need to be submitted as ds-DNA templates. As with genome samples ds-DNA templates generated from other viral types will be accepted if they meet required fragment sizes.

For certain viromes (e.g. focused on RNA viruses or ssDNA viruses, or those that have been simplified by PFGE), the Selection Committee is recommending that 1/8 of a 454 Titanium pico-titre plate be used (average 80K reads; 400nt average read length), with the option to have an additional 1/8 plate run pending the results of the first sequence dataset. In these cases, the researcher will have 2 months during which to analyze the first sequence dataset and to offer a scientific justification for additional sequencing of the same sample. If a researcher does not provide to the Broad and GBMF an affirmative justification to sequence the additional 1/8 plate, no additional sequencing will occur. For samples that are sequenced in two batches (1/8 + 1/8

plate), the six month data embargo will begin after the second round of sequencing is completed (see below for information about the data embargo).

## **2. How should I prepare my DNA?**

Isolate genomes can be concentrated and CsCl purified (see reference 1) and host genomic DNA contamination will be near 0%, but can result in low DNA yields for marine phage isolates. If DNA yields are low, a lysate extraction method can be used (see protocols) and host gDNA contamination will range from 10-50% depending upon how well the microbial fraction is removed prior to particle concentration. Sample preps should be tested for contaminating DNA using 16S/18S/recA universal primers. While the Broad will sequence samples with contaminants in them, the total coverage delivered per genome will be affected by the presence of contaminants.

Viral metagenomes Standard field collection of viruses for metagenomic sequencing (the <0.2µm size fraction) should be followed by DNase treatment and CsCl purification (see reference 1). Some have reported substantially higher DNA yield from sucrose as opposed to CsCl purification of viral concentrates. In either instance, once purified viral particles are obtained these preparations should be tested for prokaryotic DNA contamination using 16S PCR with universal primers. PCR controls should consist of 16S target added to the viral preparation (positive control) and a no-template addition (negative control). If the preparation of virus particles is found to be 16S-positive, the preparation can be 0.2µm filtered, treated with DNase and retested until no 16S signal is obtained. Once the sample is deemed free of contaminating DNA, viral DNA or RNA can be extracted from viral particles as per your favorite protocol (some example protocols are provided). See also below reference (1).

## **3. How should I quantify my DNA?**

While quantification by spectroscopy can be quick, it can also be misleadingly optimistic. This is particularly true for environmental samples that are <5ng/ul on a nanodrop. The ideal means to quantify your DNA/RNA is using picogreen/ribogreen (see protocols).

## **4. I have < 100ng of DNA, should I amplify before sending DNA to the Broad?**

It is possible to make 454 libraries from template amounts of 1ng and even less. However, reduced starting template results in an increased chance of library failure, a reduced number of sequencing reads, and the potential for bias in the sequence across the genome. Alternatively, you may wish to amplify your starting material to get more sequencing reads / data. However, each amplification method likely introduces biases (see FAQ#5). In-house, small-scale quality control of a small number of sequences is strongly recommended before submitting the materials to the Broad for full-scale sequencing.

## **5. I would like to get more sequencing reads out of my virome so will amplify my DNA before sending it to the Broad, which method should I use?**

The selection committee and the Broad do not endorse any single method over any other. However, community polling of microbial and viral researchers and a review of the literature suggests the following guidelines for your decision.

- a. MDA (multiple displacement amplification): This method under-represents the ends of linear genomes, has stochastically-determined non-repeatable biases within a genome and between genomes, and will selectively amplify ssDNA and circular genomes. The literature hints that combining multiple MDA reactions, drastically reducing reaction volumes, and including additives (e.g., trehalose) may minimize biases. It should also be noted that the Qiagen Repli-g kit is known to be contaminated with high G+C bacterial DNA from a *Delftia* spp. This contaminant DNA can be especially problematic with low starting concentrations of DNA template. Clean up from MDA reactions to remove primers will not work well with PCR clean-up kits, instead DNeasy columns are commonly used.

- b. Linker-amplification: This method was the basis of the early clone library viral metagenomes. Data suggests that linker-amplified-shotgun libraries from phage isolate genome projects have reproducible under-representation in the same regions as standard whole genome shotgun libraries regardless of cloning vector. This suggests amplification in these methods is not biased, but cloning is. Therefore linker amplification from 1-2kb fragments is likely to provide relatively unbiased amplification. (see protocols, requires Covaris S2 System to shear DNA)
- c. Sigma offers a GenomePlex Whole Genome Amplification kit that appears similar to the linker-amplification method described in 5b, but uses a proprietary shearing step that results in 400bp fragments. No data are available on it's randomness, but this might be a good options where the Covaris S2 System is not available for shearing your DNA.

### **6. What are recommendations for maximizing sequencing reads from my RNA virome?**

Purification of viral RNA from natural samples is difficult because of the relatively low abundance of RNA viruses in most samples and because even a small amount of microbial contamination can swamp the viral RNA with rRNA. If there is adequate starting material, purification in a CsCl or sucrose density gradient may be possible, otherwise purification will likely be restricted to 0.2- $\mu$ m filtration. Once purified the viruses should be treated with RNase to remove free RNA. Following extraction, the nucleic acids should be treated with DNase to remove contaminating DNA. The DNase-treated nucleic acids should be checked for remaining contamination by bacterial DNA using PCR with universal 16S primers, as outlined above. Limiting RNA can be amplified at either the RNA step (e.g., polyA tailing approach, Frias & Shi et al. 2008 though this has not been tested for viruses) or after cDNA synthesis by random priming with hexamers and reverse transcriptase where following second-strand synthesis, DNA amplification is accomplished as described in #5 (though, note that MDA, see 5a above, with cDNA often leads to heavily chimeric products and significant biases). Subsequently, PCR with universal 16S primers should again be used to check for initial contamination by rRNA. Contamination by rRNA can be very difficult to avoid; however, if contamination is found, it may be that only a minor portion of the sequences are ribosomal. If contamination by ribosomal sequences is minor, the contribution from mRNA will almost certainly be trivial. The degree of contamination with ribosomal sequences can be checked by sequencing about 20 shotgun cloned fragments. As a general rule of thumb, if <20% of the sequences are ribosomal it is reasonable to submit the sample for pyrosequencing.

### **Acknowledgments**

The Broad Institute and Moore Foundation would like to thank the Selection Committee and Matthew Sullivan for their assistance in preparing this document.

### **References**

1. Rebecca V Thurber, Matthew Haynes, Mya Breitbart, Linda Wegley & Forest Rohwer (2009) Laboratory procedures to generate viral metagenomes. *Nature Protocols* 4(4): 470-483.