SANGER SEQUENCING
WHITE PAPER

FACTORS AFFECTING SANGER SEQUENCING ROBUSTNESS AND DATA QUALITY
PATRICK WARNER, SHEA ANDERSON, ARCHANA DESHPANDE, DARYL M. GOHL, KENNETH B. BECKMAN
UNIVERSITY OF MINNESOTA GENOMICS CENTER

SUMMARY
In order to better understand the robustness of our protocols and to identify failure modes for Sanger Sequencing in the UMGC we carried out a series of experiments testing the effects of DNA template concentration, primer concentration, and sample purity on Sanger sequencing data quality. We determined that the UMGC’s Sanger sequencing processes are robust for a wide range of template and primer concentrations, and that the most likely failure modes for submitted samples center around the presence of contaminants. Finally, we compared the contaminant sensitivity profiles of the UMGC processes to two external vendors, ACGT and Genewiz, and found that the effects of template concentration, primer concentration, and contaminants were highly similar across all three providers. These experiments may help to guide researchers in preparing their samples for Sanger sequencing.

INTRODUCTION
The advent of DNA sequencing using didoxy chain-terminating nucleotides, also known as Sanger sequencing (after its inventor, Frederick Sanger), revolutionized the ability to read DNA, replacing more labor intensive and technically challenging methods. Large-scale automated Sanger sequencing underlay the effort to sequence the 3 billion base pair human genome for the Human Genome Project(1). While the work of large-scale genome sequencing has largely been usurped by Next-Generation Sequencing (NGS) technologies, Sanger sequencing remains useful for targeted sequencing of DNA from virtually any source. In particular, applications such as validating constructs, PCR amplicons, diagnostic assays, microbial identification, high throughput genotyping, and confirmation of NGS results are commonly carried out via Sanger sequencing. Sanger sequencing has a high reliability, typically achieving over 99.99% base calling accuracy.

Sanger sequencing uses a polymerase to extend off of a template-bound sequencing primer in the presence of a mixture of deoxy-nucleotide triphosphates (dNTPs) and dideoxy-NTPs (ddNTPS). As a ddNTP is incorporated into an extension product, it terminates the growing DNA molecule and labels the molecule with fluorescent tag corresponding to one of the four nucleotides. The labeled DNA fragments are then electrophoresed through a capillary and the fluorescent signal is read out to determine the sequence of bases in the template molecule (Figure 1).

Many factors can affect the quality of Sanger sequencing data. Both relative and absolute primer and template concentrations can affect the length of the read obtained, and the quality of the data. The presence of contaminants, which can arise from extraction methods, clean up procedures, and use of solutions for sample dilutions or resuspension, can also adversely affect sequencing results. In addition, intrinsic sequence properties such as GC content, secondary structure (such as hairpin-forming inverted repeats), and polynucleotide repeats (which can cause signal degradation due to slippage of the polymerase) can all cause problems with data generation and may require altered protocols.

We systematically examined several of these factors affecting Sanger sequencing data quality, in order to provide guidance on best practices for setting up and troubleshooting Sanger sequencing reactions. We tested the effect of primer and template concentration using several different test templates. In addition, we assessed the effect of spiking-in varying amounts of contaminants to samples. Finally, we examined the robustness of the UMGC’s Sanger sequencing workflow, by sequencing a set of challenging test samples at the UMGC, as well as two external Sanger sequencing providers. Using common quality metrics across these test samples, we determined that the presence of...
contaminants had the largest effect on Sanger sequencing data quality. These experiments provide insight into the common failure modes for Sanger sequencing and provide information for researchers to consider in designing and troubleshooting their Sanger sequencing experiments.

RESULTS

In order to deliver consistent high quality Sanger sequencing data, the UMGC sequences positive control samples on each Sanger plate, and carries out ongoing data quality monitoring. All Sanger plates are run with two pGEM plasmid controls, sequenced with the M13(-21) primer, and placed in asymmetrically located wells to allow for the confirmation of correct plate orientation. The pGEM control is sequenced with the same master mix cocktail and goes through the same processing as all the other samples on the plate. The pGEM control template is expected to produce 900-1000 bp of QV>20 sequence (see Appendix A for detailed explanation of Quality Values), and historical data going back four years shows that the results from this template are broadly consistent over this period (Figure 2A). Similarly, Q-scores for client samples are largely stable over time (Figure 2B). Note that in this case, there is a large range of sample types and quality (including amplicons of variable lengths, where shorter read lengths are expected), but on average, the distribution of quality scores observed is consistent.

We carried out a series of experiments in order to better understand how factors such as sample purity and template and primer concentration affect the UMGC’s Sanger sequencing process. To test the effect of template type and concentration on the UMGC’s Sanger sequencing workflow, we used three different test templates: 1) a 1.5 kb PCR product. 2) a 4.5 kb plasmid. 3) a 9.5 kb plasmid. For a 1.5 kb PCR product, the UMGC’s sample submission guidelines call for 50-100 ng of template DNA in a 6 ul sequencing reaction. We tested a range of template inputs that ranged from 25 ng-200 ng (approximately 2-fold higher or lower than this recommended range). Since plasmids are typically longer than amplicons, the recommended input mass is 50-100 ng per kb (up to a maximum of 800 ng) in a 6 ul reaction (this would correspond to 225-450 ng for the 4.5 kb plasmid and 475-800 ng for the 9.5 kb template). As with the PCR products, we tested a range of plasmid template inputs that ranged from 100-1000 ng (again, approximately 2-fold higher or lower than this recommended range for the 4.5 kb plasmid, and up to almost 5-fold lower than the recommended range for the 9.5 kb plasmid).

We sequenced each of these three templates across a range of template concentrations, and for each of these template concentration also varied the amount of primer used. For a 6 ul sequencing reaction, the recommended amount of primer is 3.2 pmol. We tested a range of primer quantities from 0.8 pmol (4-fold lower than recommended) to 25.6 pmol (8-fold higher than recommended). We used multiple criteria to evaluate sequence quality. For consistency we use the number of QV >20 bases as a metric in this document, but similar results were seen with metrics such as average QV, or the position of the last base in a read above a QV threshold of 30, 40, or 50.
The effect of template and primer concentration on sequence quality was minimal within the matrix of conditions that we investigated. For the 1.5 kb plasmid (Figure 3A) and the 4.5 kb plasmid (data not shown), there was almost no discernable effect of suboptimal primer or template concentration. For the 9.5 kb plasmid (Figure 3B), a moderate decrease in quality was seen with excessive template quantity at the high and low extremes of primer concentration, suggesting that optimized primer concentration may become more critical in the case of template DNA that exceeds the recommended input amount.

Next, we assessed the effect of contaminants on Sanger sequencing data quality. We chose a battery of common contaminants that may be present in a sample either due to the use of inappropriate buffers (such as TE, which contains EDTA, a well known inhibitor of the enzymes used in Sanger sequencing), or through carry-over from the DNA extraction process (such as SDS, which is commonly used in cell lysis; phenol:chloroform, which is used in extraction of nucleic acids; excess dNTPs, which are used in PCR; as well as various salts and alcohols, which are used in the precipitation of DNA). We spiked a range of concentrations of these inhibitors into a Sanger reaction with optimal input amounts of the 4.5 kb plasmid and primer tested above. We found that, in contrast to the effects of varying template and primer inputs, the presence of some of these contaminants had a dramatic effect on sequence quality. In particular, EDTA, SDS, and NaCl all strongly inhibited the Sanger sequencing reaction (Figure 3C).

In order to determine whether the sensitivity to these contaminants reflected intrinsic limitations of the Sanger sequencing process, or a lack of robustness in the UMGC’s Sanger sequencing workflow, we prepared replicate plates containing the above template and primer concentration tests, along with an expanded set of contaminants. We sequenced this plate in the UMGC and also sent copies to two external providers of Sanger sequencing (ACGT and Genewiz). For both the template/primer concentration sweeps and the contaminant profiles, the data from all three sites was highly similar. By and large, inhibition was observed with the same contaminants at roughly the same concentrations across all three facilities (Figure 4). One exception to this was the 1 mM EDTA condition (notably, this is the concentration of EDTA in TE buffer), which inhibited the UMGC’s and Genewiz’s sequencing reactions, but had only minimal effect on the ACGT reaction (Figure 4C). For other contaminants, data quality from ACGT was also subtly improved relative to the UMGC and Genewiz (Figure 4E, 4F).

DISCUSSION

The data presented here demonstrate that the UMGC’s sequencing protocol is robust and stable over time (Figure 2), and is tolerant to a range of template and primer concentrations (Figure 3). In addition, we explored the effect of contaminant spike-ins on our Sanger sequencing process. We found that a number of commonly used DNA extraction reagents and buffer components could negatively affect Sanger sequencing data quality. In order to determine how the robustness of the UMGC’s processes compared to that of other Sanger sequencing providers, we sent the same test plate to two external vendors. The contaminant sensitivity profiles were largely similar among the three laboratories (Figure 4), indicating that these contaminants are affecting intrinsic aspects of the Sanger sequencing process.

These experiments shed light on the common failure...
modes for Sanger sequencing reactions. While suboptimal template and primer concentration do affect the quality of Sanger sequencing data(2), particularly for templates that are difficult to sequence due to secondary structure, GC content, or length, we found that for 1.5-9.5 kb PCR and plasmid templates, our workflow (as well as that of two external vendors) is tolerant of deviations of 2-5 fold from the recommended template concentration and between 4-8 fold from the recommended primer concentration.

In contrast, the presence of contaminants, such as EDTA, SDS, phenol:chloroform, salts, and to a lesser extent dNTPs had strong concentration-dependent effects on Sanger sequencing data quality. These data suggest that when troubleshooting a failed Sanger sequencing reaction, assessing whether these contaminants, many of which are commonly used reagents in the processes of DNA extraction, amplification, and elution/resuspension, should be a first step in the troubleshooting process.

METHODS

OUTLINE OF THE UMGC SANGER SEQUENCING PROCESS.

All Sanger plates in the UMGC are run with two pGEM plasmid controls, placed in asymmetric wells for corroboration of plate orientation. The pGEM 3Zf(+) is sequenced with the M13(-21) primer (5’TGT AAA ACG ACG GCC AGT-3’). The pGEM control is added to the plates at the same time, in the same manner, and the same volume as the samples. The pGEM controls have the same cocktail added as the samples then go through the same processing as all the other samples on the plate.

When all samples and controls are transferred to the reaction plate, BigDye cocktail is added. The plate is then sealed, vortexed and centrifuged before starting PCR cycling in an ABI Fast Thermocycler. During PCR the template undergoes denaturation, annealing, extension to incorporate the fluorescently labeled ddNTPs.

After the cycling of PCR on the thermocycler the plate is cleaned up using Magbio HighPrep DTR beads automation with a Beckman NXp robot. The clean up process will removes most or all of remaining contaminants, primers, BigDye, and ethanol from the reaction wells. The cleaned up template is transferred to a fresh plate for sequencing and is loaded onto the 3730xI DNA Sequencer for analysis.

PRIMER AND TEMPLATE CONCENTRATION TESTS

We tested three differently sized templates and their corresponding sequencing primers with a matrix of template and primer concentrations. These templates were chosen on the basis of not having problematic GC content, secondary structure, and polynucleotide or repeat regions, so that we could better assess the effects of the other experimental variables under investigation. The primers used were within the recommended specifications for Sanger sequencing.

The 1.5 kb PCR product (GC content = 49.9%), was a full-length Campylobacter jejuni 16S rRNA gene amplicon, and was run with varying template inputs of 25, 50, 100, 200 ng versus the varying primer concentrations of 0.8, 1.6, 3.2, 6.4, 12.8, to 25.6 pmol. This amplicon template was prepared with a mixture of the following four forward primers and the following reverse primer:
27F_YM_1_Nextera: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGA
27F_YM_2_Nextera: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGG
27F_YM_3_Nextera: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGG
27F_YM_4_Nextera: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGG
1492R_Nextera: GTCTCGTGTTAGCTGAGATCTACACATAGAGAAGAGATTTTGATCCTGGCTCAG
THE FOLLOWING SEQUENCING PRIMER WAS USED TO SEQUENCE THIS AMPLICON:
p5: AATGATACGGCGACGCACAGAGATTTTGATCCTGGCTCAG

We also tested two plasmids of differing lengths, both of which were previously cloned by UMGC staff. The plasmids and their respective sequencing primers were as follows:

pTOPO-01_ABa_V4_TCT_A (4.5 kb plasmid: GC content = 52.5%)
Sequencing primer: M13F: TGTTAAACGACGAGATCTACACATAGAGAAGAGATTTTGATCCTGGCTCAG

pBPHLWL-LexA (9.5 kb plasmid: GC content = 49.2%)
Sequencing primer: pPromF(seq): CAGTGCACAGGTGGTTTGCTTGTTGA

Plasmid products were sequenced with varying template inputs of 100, 400, 700, 1000 ng added to the varying primer concentrations of 0.8, 1.6, 3.2, 6.4, 12.8, to 25.6 pmol.
CONTAMINANT SPIKE-IN TESTS

We prepared test samples for contaminant testing using the 4.5 kb pTOPO-01_ABa_V4_TCT_A plasmid and its corresponding sequencing primer with optimal template and primer concentrations; 400 ng template and 6.4 pmol primer. The following contaminants were then spiked in at the indicated concentrations:

- Isopropanol: 0.625, 1.25, 2.5, 5, 10%
- EtOH: 0.625, 1.25, 2.5, 5, 10%
- Phenol:chloroform: 0.625, 1.25, 2.5, 5, 10%
- BSA: 0.625, 1.25, 2.5, 5, 10%
- EDTA: 0.0625, 1.25, 0.25, 0.5, 1 mM
- NaCl: 0.0625, 1.25, 0.25, 0.5, 1 mM
- dNTP: 0.0625, 1.25, 0.25, 0.5, 1 mM
- SDS: 0.015625, 0.03125, 0.0625, 0.125, 0.25%

We made a single master test plate containing both the template and primer concentration sweep experiments and the contaminant spike-in experiments, and made replica plates from this master plate for testing at the three facilities. The sample plate was submitted to UMGC blinded as normal client samples. For the submissions to Genewiz and ACGT, we followed the sample submission volumes listed on the vendor websites.

APPENDIX A

The QV score is a per-base estimate of the base caller accuracy. It is calculated as $QV = -10 \log_{10}(Pe)$ where $Pe$ is the probability of error. The base caller generates QVs from 1 to 99, with 1 being low confidence and 99 being high confidence.

**TABLE 1**

<table>
<thead>
<tr>
<th>Quality Value (QV)</th>
<th>Errors per 1000 bases</th>
<th>Probability of Error (Pe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>0.1</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td>40</td>
<td>0.1</td>
<td>0.0001</td>
</tr>
<tr>
<td>50</td>
<td>0.01</td>
<td>0.00001</td>
</tr>
</tbody>
</table>

REFERENCES: