Pilot study: Surveying Complex Samples for Synthetic Elements by Targeted Enrichment

Author: Andrew Bergman

Supervisor: Andreas Sjödin

LUNDS UNIVERSITET

Contents

Abstract

Our ability to genetically engineer microorganisms is growing rapidly. We can perform small, precise modifications using techniques such as CRISPR-Cas9 to whole-virome synthesis and/or large genetic insertions. These advances reduces the threshold for synthetic biology to be employed maliciously. Preparations against pandemics or deliberate biological attacks currently rely on taxonomic identification of microorganisms and inference of their pathogenicity. In the modern landscape of bioengineering, taxonomic identification may be misleading as commonly benign microorganisms may be altered to increase their virulence and pathogenicity. Since modern bioengineering can fail to detect harmful microbes, there is a gap in the current methodology. In order to address this gap, this project aims to assess if in-solution hybridization (targeted enrichment) is a feasible approach to identify microorganisms by selectively enriching DNA elements that could indicate synthetic intervention. The approach utilizes a custom bait panel that targets plasmids and antibiotic resistance genes (ARGs). The bait panel was applied to a fecal sample set and a spike-in sample set (where a cell line, VERO E6, is mixed with a concoction of plasmid sequences). The enrichment outcome is preliminarily successful for initial DNA concentrations as low as 2.5e-5 ng. Background depletion works differently well depending on sample set, one of them yielding background sequence reduction from 90% to 10%. Continued experimentation should build upon the exploratory results achieved in this study by validating the results and continuing the optimization of the bait panel.

1. Introduction

Biological weapons are composed of living matter that causes damage or death to animals and/or plants. Considered weapons of mass destruction, they include pathogens such as *Bacillus anthracis* and *Yersinia pestis* and toxins such as the *Botulinum toxin*. Toxins can be considered either chemical or biological weapons as per the Bioloigcal Weapons Convention and the Chemical Weapons Convention, respectively (Feakes D, 2017).

Long before humans' ability to genetically engineer pathogens, biological weaponry has been used against humans, livestock and crops. Native American and African tribes were early adopters of biological weapons, having been known to dip arrowheads into poison to more effectively dispose of adversaries (Madsen et al., 2006). Another well-known instance of biological weapons usage occurred in 1763 when Europeans purposefully infected blankets with the variola virus (smallpox) and subsequently gifted/sold them to native populations in North America (Christopher et al., 1997). Malicious use of biological agents continued during World War 1 as Germany initiated the first state-funded bioweapons program. The aim of the program was to infect livestock from various countries with *B. anthracis* and *Pseudomonas mallei (Riedel S., 2004)*. During World War 2, both the germans and japanese conducted bioweapons testing on prisoners, Japan alone causing more than 10000 casualties (Venkatesh & Memish, 2003). There are additional modern examples of biological weapons usage; During the gulf war (1991), anthrax was successfully released (Venkatesh & Memish, 2003) and in 1993 the Aum Shinrikyo cult attempted to release anthrax in Tokyo nine times before successfully releasing serine gas in 1995, killing 12 and injuring more than 5000 people (Pletcher K, 2010).

The rapid development of AI, robotics and additive manufacturing are three dual use research concerns regarding biological weapons. These technological advances have great potential benefits, however, in the wrong hands, they can speed up development and deployment of synthetic biological agents. Applying artificial intelligence and machine-learning to optimize virulence attributes of microorganisms is feasible, so is the automation of previously manual labor using robotics (Brockmann et al., 2019). Several toxins have pharmaceutical or cosmetic utility. An unfortunate consequence of the commonplace utility is that the toxins can be produced in the shadow of legitimate industries, potentially placing biological agents within reach for malignant actors. For instance, the botulinum toxin is the most lethal toxin known to man (LD50 = 3 ng/kg), and is at the same time utilized cosmetically to mitigate skin wrinkling (Padda & Tadi, 2023) (Witmanowski H, & Błochowiak K., 2019) and pharmaceutically to combat hepatic and renal impairment (Padda & Tadi, 2023) (Brockmann et al., 2019).

Biodefense emerged as a well-funded field of research after the anthrax mail attack in the United States, 2001. The attack signified that bioterrorism no longer is theoretical and needs to be prepared for (Tegos, 2013). The effort to increase biosecurity was further increased after the SARS-CoV-2 pandemic of 2019. The pandemic opened the eyes of the public with regards to the damage a biological agent can achieve, both physically to individuals and systemically to societies. Societal damages include reduced economic output (Arias et al., 2022) and subsequent mental illness following societal lockdowns (Penninx et al., 2022).

Current biosurveillance methods rely on the notion that pathogenicity and virulence is strongly predicted by taxonomy. Thus, the methods rely on microbial classification and PCR techniques as well as brute-force metagenomics. The idea of broad-scope biosensors have been developed (Tegos, 2013), however they are not sufficiently sensitive to detect genetic manipulation. Microbial classification is not necessarily straightforward, for instance, *B. Anthracis* and *B. Cereus* are hard to distinguish from each other due to the sole difference between the strains being that *B. Anthracis* carries two unique plasmids that *B. Cereus* does not (Tegos, 2013). A lot of biodefense resources are allocated to understanding risk species (Tegos, 2013), this however, might not prove fruitful when facing a meticulously engineered, commonly benign microorganism. There is ongoing research focusing on identifying markers of genetic engineering, one such approach is Synsor (Tay et al., 2024). There are also several alignment-based methods that require reference genomes (Wu et al., 2022). However, when dealing with novel biological weapons, this information is likely unavailable. Another approach that has been trialed is the use of artificial neural networks (ANNs) trained on the k-mer signatures of natural and synthetic plasmids (Tay et al., 2024). This approach yielded some ability to detect genetic engineering in samples, however the sensitivity and specificity is low (Tay et al., 2024).

This project is a pilot study inspired by Intelligence Advanced Research Project Agency's (IARPA) attempt to create a biosurveillance method. IARPA's final results are not available to the public, their method has only been discussed briefly. This study describes the development and assessment of an in-solution hybridisation bait panel, targeted toward plasmids and ARGs. In-solution hybridization/Targeted enrichment is a method of selecting specific sequences and amplifying them. This method enables the capture of low abundance DNA, which would otherwise be absent using conventional NGS techniques. By enriching sequences of concern, they can be captured with more sequencing depth and with higher accuracy. We hypothesize that the captured sequences contain flanking regions that can be used for sequence classification, furthermore, all sequences will be genetically annotated. Using this approach, we aim to get an understanding of the enrichment profiles on different sample sets. In extension, this project also aims to be the first stepping-stone in developing a modern biological weapons surveillance framework (figure 1).

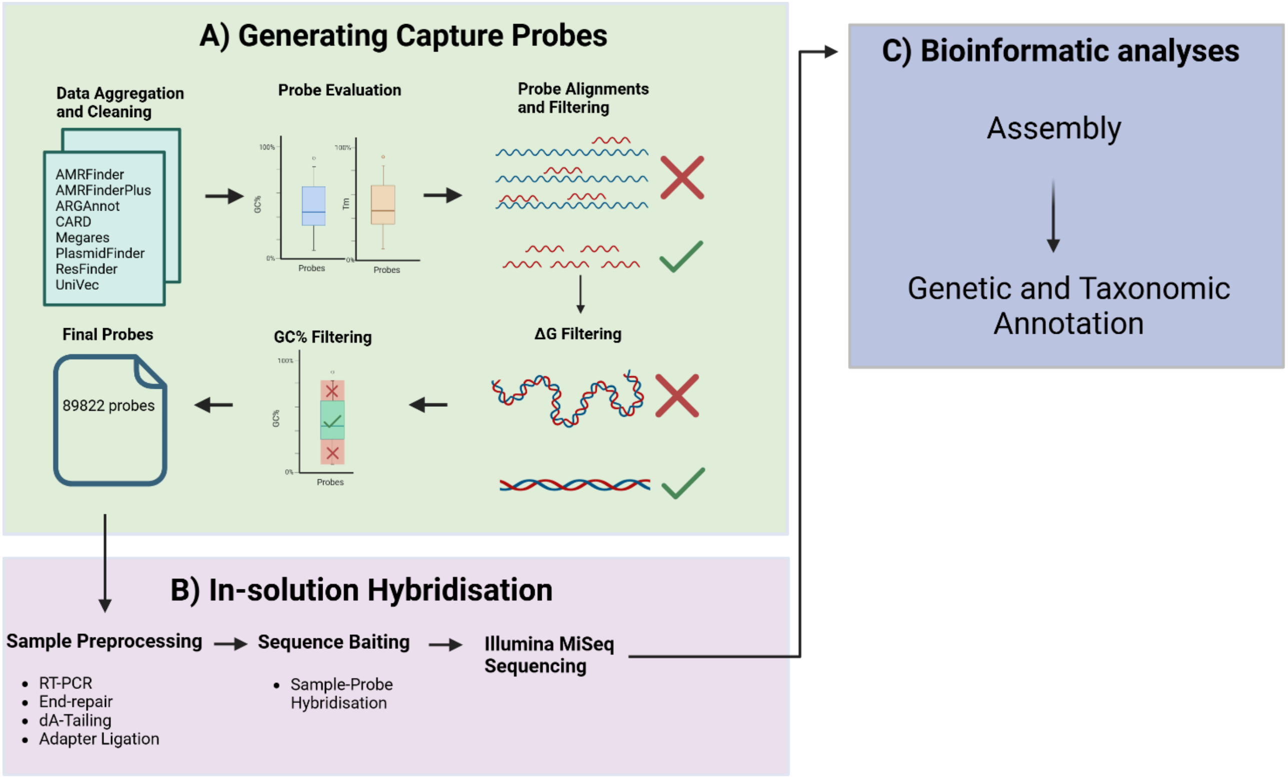

Figure 1. The workflow of this study entails generating bait sequences, performing in-solution hybridization, sequencing and bioinformatic analyses. **A)** Data is aggregated from databases containing ARGs and plasmid sequences. Syotti is employed to generate baits from the database. The baits are subsequently filtered based on alignment to *black-list* sequences and their chemical and molecular properties (Gibbs free energy, GC% and Tm). **B)** After the bait filtering, in-solution hybridization (targeted enrichment) follows. **C)** The captured sequences are assembled into contigs which are subsequently classified and genetically annotated.

2. Methods

2.1 Generating in-solution hybridisation baits

DNA sequences corresponding to ARGs were retrieved from AMRFinder (Feldgarden et al., 2020), AMRFinderPlus (Feldgarden et al., 2021), ARGAnnot (Gupta et al., 2013), CARD (Alcock et al., 2023), megares (Doster et al., 2019) and ResFinder (Florensa et al., 2022). Plasmid sequences were retrieved from plasmidfinder (https://github.com/kcri-tz/plasmidfinder), NCBI's UniVec and NEB's plasmid backbones. The sequence data was cleaned from unwanted characters and any lowercase nucleotides were converted to their uppercase counterparts. The data was followingly concatenated and Syotti was employed to generate 120-mer bait sequences with varying numbers of mismatches. After having generated sets of bait sequences covering the input database, the GC-contents and melting temperatures were evaluated using *gc_tm_boxplot.R*. Followingly, the sequences were filtered to ensure non-complementarity to mitochondrial and chloroplast gene sequences (MMSeqs2 was used for alignment). Any bait aligning with >80% identity to these genes were filtered out from the bait set (*remove_problematic_baits.R)*. Similar filtering was subsequently conducted against the human genome (GCF_000001405.26), any bait aligning with >90% identity to the human genome were removed *(filter_problematic_baits.py, remove_problematic_baits.py)*. The remaining baits were filtered on the basis of GC-content so that the remaining bait sequences had 35<GC%<75 *(filter_gc.R)*. Gibbs free energy was calculated for the remaining baits using viennaRNA's RNAfold. The output was parsed using *filter_free_energy.R,* any bait sequence hosting Gibbs free energy less than -40 were removed using *remove_self_compl.R*. Unfiltered, filtered and semi-filtered baits were compared in terms of GC-content and melting temperature to assess the effect of filtering *(bait_filter_comparison.R)*. Finally, the bait coverage of the original datasets was assessed using *generate_overlap_plot.R*. The full bait generation Snakefile is located at https://github.com/AndrewBergman1/Targeted_enrichment. The filtering thresholds (GC-content, alignments, ΔG) could be more stringent, however, this would result in fewer probes.

2.2 Sample sets

Two sample sets were used to evaluate the bait panel. *Sample set A* is composed of a dilution-series of a fecal reference sample (D6323+) provided by Zymo Research (Zymo Research, 2021). *Sample set B* is a green monkey kidney cell line (VERO E6), spiked by a 20 plasmids commonly applied during bioengineering (The concentrations of each individual plasmid is unknown.) (S1).

2.3 In-Solution Capture Reaction

The in-solution hybridization workflow entails library preparation, bait hybridization, PCR amplification and Illumina sequencing. In-solution hybridization enables the capture of low-abundance reads that would otherwise be drowned using conventional NGS approaches. Using a custom bait panel targeting antibiotic resistance genes and plasmid sequences, those are selectively amplified.

Below are brief descriptions of each step in the in-solution hybridization protocol (https://github.com/AndrewBergman1/Targeted_enrichment/tree/main/targeted_enrichment), authored by Olivia Wesula Luande.

2.3.1 Sample preprocessing

The DNA is enzymatically fragmented, then end-repaired and dA'd, prohibiting any unwanted ligation as per *DNA fragmentation, end repair and dA-Tailing*. TWIST universal adapters are enzymatically ligated to the ends of each amplicon (*Ligation, TWIST universal adapters)* and the adapter-ligated DNA sequences are purified. The purpose of the adapter ligation is to facilitate amplification of the sequences using qPCR. The amplification utilizes TWIST UDI primers that are complementary to the TWIST universal adapters, leading to amplification of all bounded DNA. The DNA is further purified from residual PCR reagents, and from here on, the samples are ready for baiting (*PCR amplification using TWIST UDI primers, Purification and QC*).

Zymo samples (sample set A) were analyzed as a dilution series, where the highest dilution was the stock solution received from ZYMO (25ng), the subsequent samples were diluted by log10 from the initial concentration (1x) to 1e-6x. The plasmid spike-in samples (sample set B) were analyzed using a smaller range between the highest and lowest dilutions; importantly, sample set B includes a negative and positive control. The positive control is unenriched, containing 25 ng DNA, of which 2.5 ng is plasmid DNA (concoction of 20 plasmids). The negative control is the background cell line (VERO E6), without any plasmid sequences.

2.3.2 Sequence baiting

After having conducted sample pre-processing (as described in section 2.3.1), the samples are mixed and hybridised with the bait sequences. The hybridization entails a 16 hour qPCR during which the baited sequences are extended (*Hybridise captured baits with pools*). The enriched sequences are extracted from the samples using Streptavidin beads (*Bind Hybridiszed targets to Streptavidin beads)* and then purified (*Purification 3*). The captured sequences are followingly amplified via qPCR (*Post-capture PCR amplification, purifiication and QC*), and then purified from residual PCR reagents (*Purification 4*). After this final purification, the target-enriched DNA sequences are extracted.

2.4 DNA sequencing

The enriched DNA library was sequenced using Illumina Miseq. The MiSeq was conducted using 300x2 cycles, pair-end sequencing.

2.5 Assembly

The captured reads were assembled using Trinity, an assembler commonly used for RNA-seq. Trinity works by extending reads, focusing on capturing and clustering isoforms. Trinity is favored to other assemblers because it emphasizes the assembly of different isoforms. This is relevant when assembling plasmid sequences with common vector backbones, where they can be distinguished as different isoforms.

2.6 Sequence alignments

Many-against-Many searching 2 (MMSeqs2) is designed for handling large amounts of data. MMseqs2's sensitivity matches the sensitivity of PSI-BLAST and runs up to 400 times faster, making it highly scalable compared to conventional BLAST (Steinegger et al., 2017). The mmseqs2 algorithm entails prefiltering and alignment. The prefilter module performs k-mer matching between the query and target database, the alignment module performs vectorized Smith-Waterman alignments.

2.7 Sequence classifications

Kraken2 yields fast taxonomic classification utilizing k-mer indexing. Kraken2 was employed with the standard database (532 GB) using the confidence threshold of 0.1, meaning that 10% of the k-mers of any given read shall map to a phyla in order to classify the read as such. The Kraken2 snakerule is found at https://github.com/AndrewBergman1/Targeted_enrichment.

2.8 Assessing microbial abundance

Sequence classification was conducted using Kraken2. The classified phyla were extracted and the undiluted sample and control sample and were normalized against in sample set A and B, respectively. Any phyla present in the sample sets and absent in the control/undiluted samples were initialized to 1e-3, from which the microbial abundance fold change was calculated. See *spike_in_stats*.R and *fecal_stats*.R on https://github.com/AndrewBergman1/Targeted_enrichment.

2.9 Assessing genetic diversity

Contigs were aligned to NT, allowing for five hits per query. The fraction of plasmid/ARG hits was calculated per sample, as well as the representation of unique plasmids/ARGs per sample. See *enrichment_analysis*.R on https://github.com/AndrewBergman1/Targeted_enrichment.

3. Materials

3.1 Equipment

Instrument	Reagent(s)	Software	Purpose
Invitrogen Qubit 3.0 Fluorometer	1x dsDNA HS Working Solution	v.3.0	DNA concentration estimations
Advanced Analytical Fragment Analyser 3117	Agilent NGS fragment (1-6000bp), 500, 4C	v.1.2.0.11	Analysing NGS library fragment lengths
MiSeq (Illumina)	MiSeq reagent kit v3 600 cycles	v. 4.1.0.656	Sequencing enriched samples

3.2 Samples

Sample	Constitution
Fecal sample set | sample set A	ZYMO 6323+
Spike-in sample set | sample set B	VERO E6 (Monkey) spiked with 20 plasmids at varying concentrations (S1).

3.3 Software and scripts

Scripts, mamba environment files and snakefiles are all available at

https://github.com/AndrewBergman1/Targeted_enrichment.

4. Results

4.1 Bait generation

Syotti was used for bait generation, producing 120-mer baits with 0, 10, 20 and 40 allowed mismatches with 100% database coverage. Syotti generated fewer baits when permitting more mismatches (table 1). The GC-content and melting temperatures were evaluated for each bait set, but no differences were observed (S1). As multiple filtering steps follow, the bait set with zero mismatches was chosen to proceed with. After removing 40307 baits that were complementary to chloroplast- and mitochondrial sequences as well as the human genome, 100566 baits remained. The self-complementarity of each bait was assessed, a threshold was set at delta-G<=-40, leaving 89822 baits after filtering (figure 3A). The threshold could be more stringent, but that results in fewer baits. Finally, the baits were filtered on the basis of GC-content, retaining baits that host GC-content between 35% and 75%. Comparing the unfiltered, semi-filtered (alignment-filtered) and fully filtered bait panels, the IQR (GC-content) is reduced from 0.158 to 0.13 and the IQR (Tm) is reduced from 6.49 to 5.8 (figure 3C). Shannon's sequence entropy was calculated for the filtered bait panel, the mean entropy is 1.95 bits (IQR: 0.0474), median entropy is 1.06 bits, minimum entropy is 0.92 bits and the maximum entropy is 2 bits (figure 3D).

Table 1: Sequence coverage and number of generated baits at different numbers of permitted nucleotide mismatches during bait generation.

Allowed Mismatches	**Sequence Coverage %**	**Number of baits**
0 (A)	99.98	140873
10 (B)	99.99	111707
20 (C)	99.99	103129
40 (D)	99.99	92971

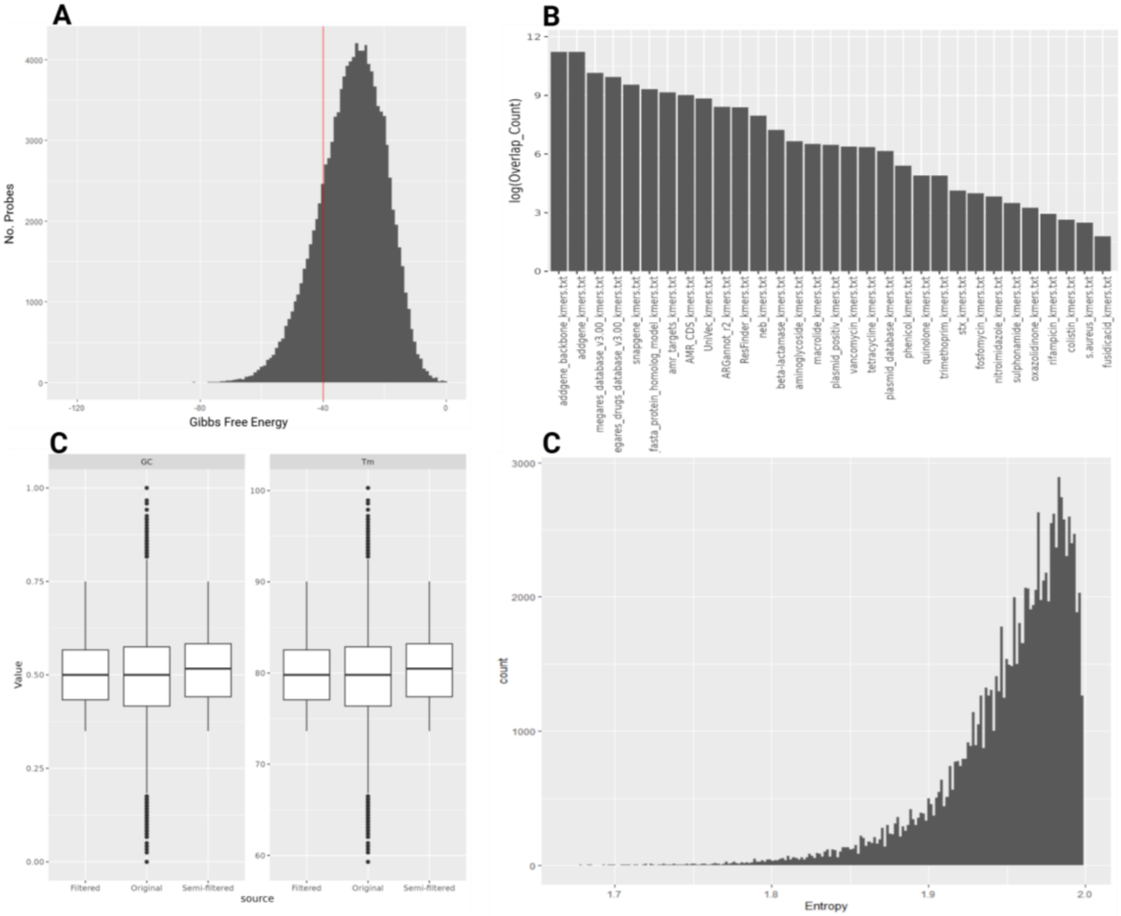

Figure 2. A) Distribution of Gibbs free energy (delta-G) throughout the semi-filtered bait sequences. Sequences hosting ΔG > -40 were retained from the semi-filtered baits. **B)** Final bait sequences' overlap with the initial datasets. **C)** Comparison of original, semi-filtered and filtered data in terms of GC-content and melting temperature. There are 140940 original baits, 100566 semi-filtered baits and 89822 filtered baits. **D)** Distribution of Shannon entropy throughout the bait sequences in the filtered bait set.

4.2 Sequence origins

The enriched sequences were assembled into contigs which were then classified to phyla. The purpose of sequence classification is to assess if the enrichment selects for any particular phyla in the different sample sets and it contextualizes the subsequent genetic annotation. Sample sets with few phyla may not have a high sequence variance compared to samples with diverse phyla. The different sample sets hosts different compositions of domains. Sample set A contain between 0.06-6% human DNA, 93-100% bacterial DNA and 0.04-0.5% viral DNA (table SVII). Sample set B host between 0.8-79% human DNA, 17-84% bacterial DNA and 0-14% viral DNA (table SVI). Observing the microbial abundance, sample set B's control sample hosts a low diversity of phyla. Upon enrichment, the diversity of the samples increases. The microbial composition remains stable in sample B onwards (0.25e-2 ng to 10 ng) (figure 3A). Commonly enriched phyla along sample set B include Bacteroidota and Kitrinoviricota. The remaining phyla are absent in at least one sample. Atribbacterota and Peploviricota are outliers as they are only present in one sample each, sample B (0.05ng) and G (25ng) respectively. The reason for why most phyla are absent in at least one sample per sample set may be stochastic or due to an overabundance of false positives from the classification, which is a feasible outcome when classifying with low confidence (confidence: 10%). In sample set A, the undiluted samples are enriched. The microbial abundance profiles of the most diluted samples (samples F-H) are similar to the microbial profile of the undiluted samples (samples A-B). The intermediary dilutions (C-E) are outliers (figure 3B).

The relative microbial abundance in each sample was assessed by normalizing to either the 1x dilution sample or the control sample, for sample set A and B, respectively. Sample set B hosts a wide array of novel microorganisms that are absent in the control sample, these phylas' fold changes are calculated by initializing their abundance at 1e-3 (figure 4). Kitrinoviricota is highly enriched in samples C-F (0.01-10ng), the highest fold change achieved in sample D (1ng), being enriched 364.8 times the control value. Bacterioidota is another highly enriched phylum, having been enriched 200-268 times the control sample throughout the sample set (figure 4). The most enriched phylum in sample set A is Cossaviricota in sample E (10e-3 ng), having been enriched by 855.5 times. Kitrinoviricota is another highly enriched phyla, being 296 times more abundant in sample D than in the undiluted samples (A-B) (figure 5). The largest overall microbial fold change was observed in sample E from sample set A and sample H (0.025ng) from sample set B at 60.5x (table II) and 59.9x (table III), respectively. The depletion of background sequences and the amplification of select phyla can be observed in sample H in sample set A. The overall fold change between sample H and the undiluted samples (A-B) is 0.85x, however, several phyla are distinctly enriched and others distinctly diminished (figure 5F).

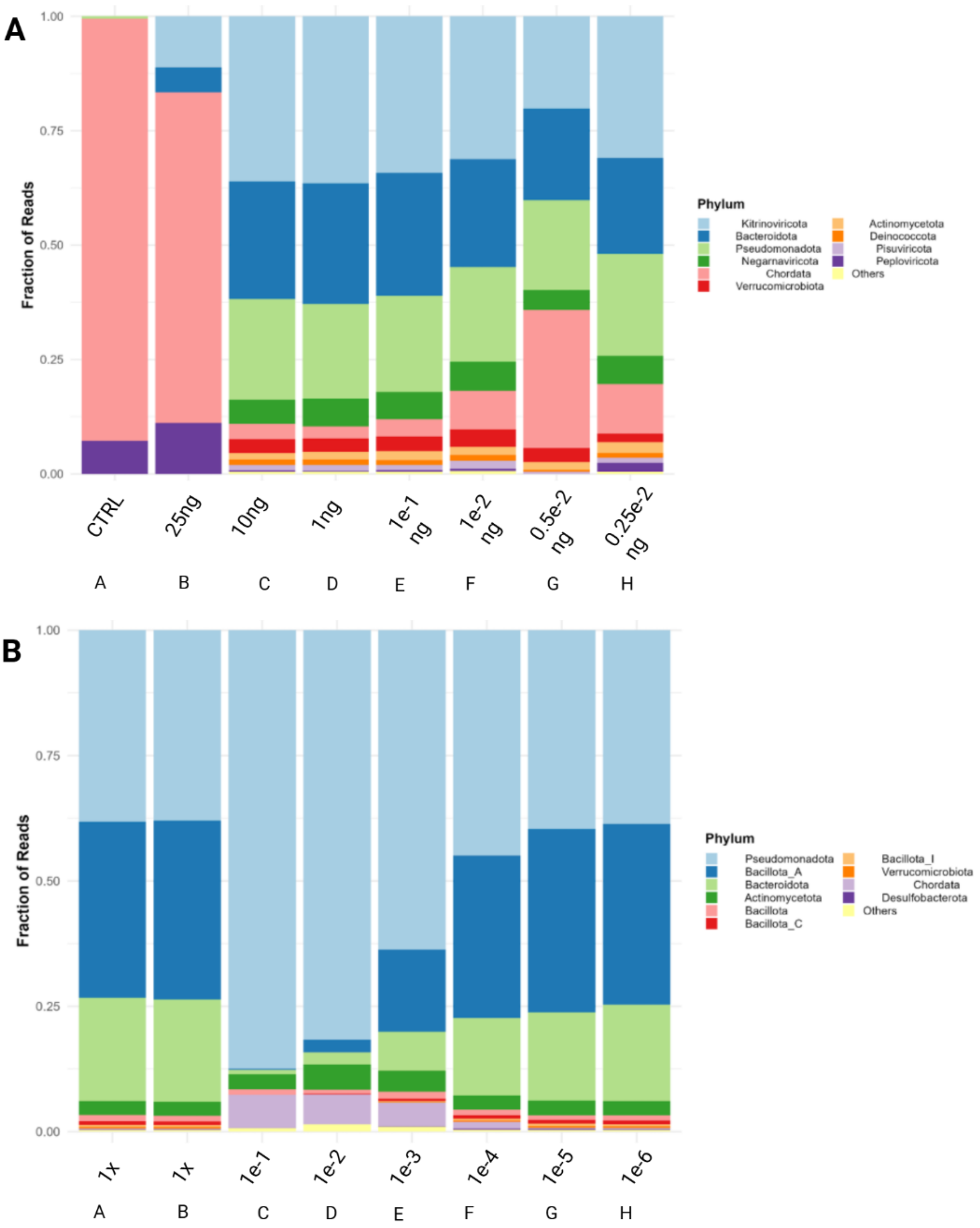

Figure 3. Microbial abundance in the sample sets used to evaluate the bait panel. 10 phyla were permitted per plot, remaining phyla are grouped together as "others". **A)** Sample set B, the spike-in sample set (VERO E6 + 20 plasmids). **B)** Sample set A, The fecal sample set (Zymo 6323+).

Table 2. Average microbial fold-change (enrichment) in sample set A. The enrichment is calculated by normalizing the microbial abundance of each sample against the undiluted (enriched) sample and then averaging the fold-change for all microbes.

Sample	**Average enrichment**
10e-1	23x
10e-2	45.6x
10e-3	60.5x
10e-4	3.26x
10e-5	0.88x
10e-6	0.85x

Table 3. Average microbial fold-change (enrichment) in sample set B. The enrichment is calculated by normalizing the microbial abundance of each sample against the control sample and then averaging the fold-change for all microbes.

Sample	**Enrichment**
25 ng	42.2x
10ng	23.8x
1ng	36.5x
0.1ng	37.5x
0.05ng	57.1x
0.025ng	59.9x
0.01ng	57.7x

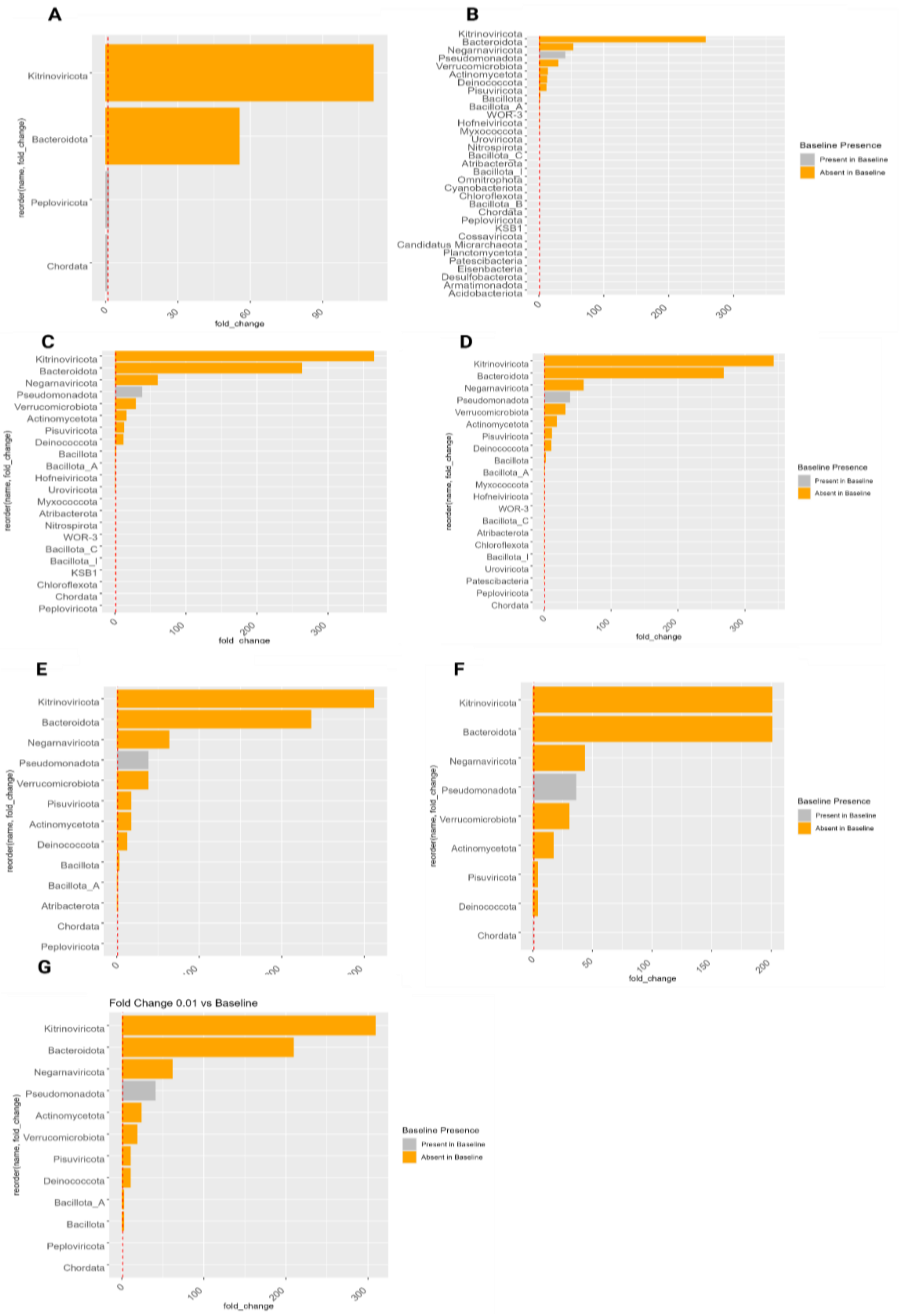

Figure 4. Microbial distribution and enrichment in sample set B. A) Sample A (25 ng) B) Sample B (10 ng) C) Sample C (1 ng) D) Sample D (0.1 ng) E) Sample E (0.05 ng) F) Sample F (0.025 ng) G) Sample G (0.01 ng). The jagged red line corresponds to the control samples' abundance of the microorganisms (fold change = 1). Orange bars represent microorganisms that are absent in the control sample, for those microbes, a reference baseline value of 1e-3 is attributed to calculate the fold change.

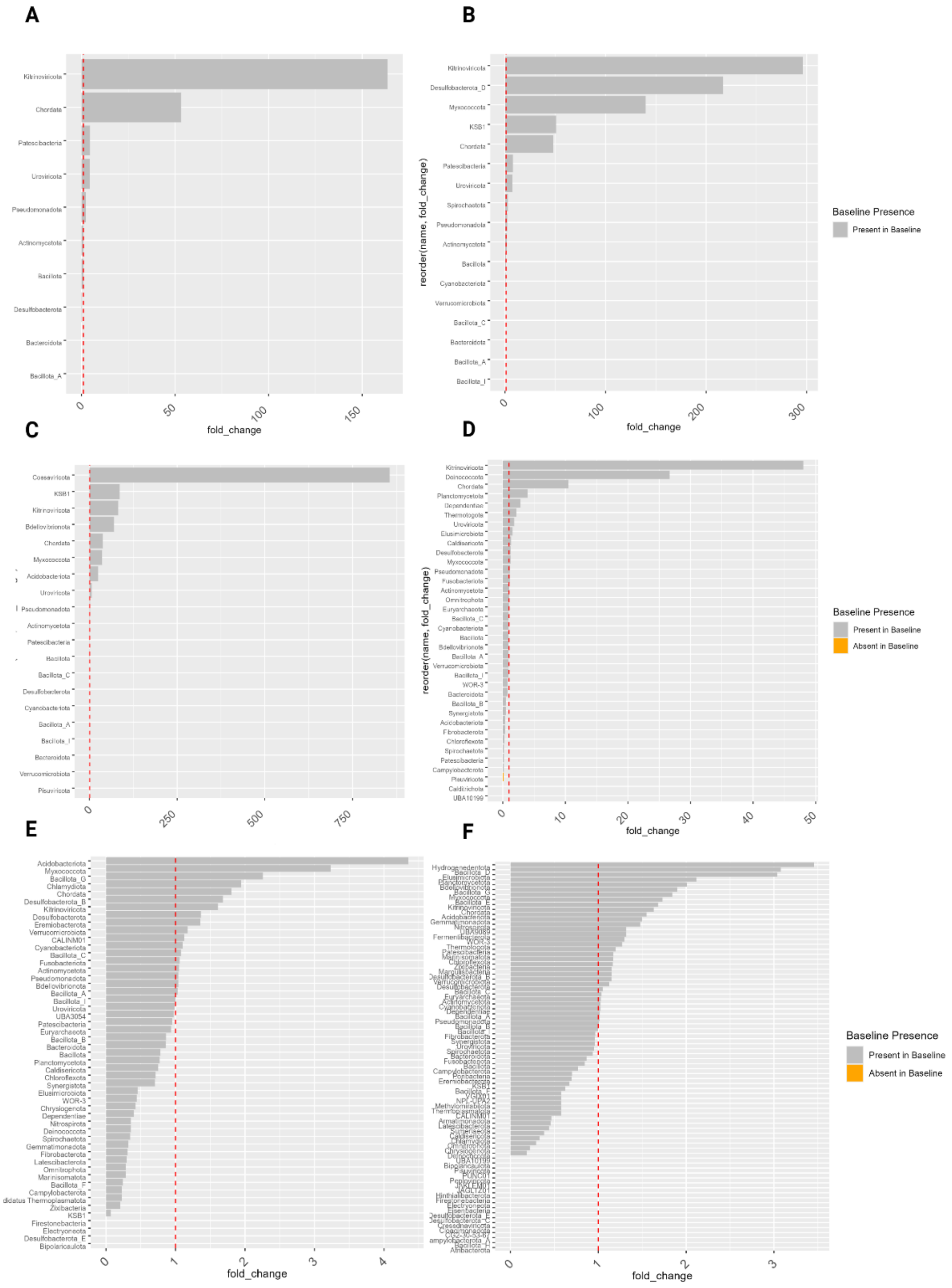

Figure 5. Fold change in sample set A A) Sample C (0.1x dilution) B) Sample D (0.01x dilution) C) Sample E (0.001x dilution) D) Sample F (1e-4 dilution) E) Sample G (1e-5 dilution) F) Sample H (1e-6 dilution). The jagged red line corresponds to the control samples' abundance of the microorganisms (fold change = 1). Orange bars represent microorganisms that are absent in the control sample, for those microbes, a reference baseline value of 1e-3 is attributed to calculate the fold change.

4.3 Genetic diversity

The genetic diversity is assessed by first assembling the reads into contigs and then aligning them to NT using megaBLAST. Five alignments were permitted per query, and by parsing the names of the alignments, they are classified as either plasmid/vector, ARG or other. The highest post-enrichment DNA concentration in sample set A was observed in sample C (1.58 ng/ul) (figure 6B). In sample set B, sample B yielded the highest post-enrichment DNA concentration (5.22 ng/ul) (figure 6A). In sample set A, the highest extent of unique plasmids and ARGs were observed in sample G, the most diluted sample, hosting 55.8% of all unique plasmids and 51.6% of unique ARGs found throughout the sample set. At the same time, the sample only hosts a small fraction of plasmids/ARGs compared to other sequences (5.1% ARGs, 4% plasmids, 91% other) (Figure 6A, C). The pattern of small number of ARGs/plasmids representing a large fraction of the overall diversity is present throughout sample set A (figure 6C). The ARG prevalence remains proportionally the same throughout the dilutions, ranging from 4.37-5.58%. The range of plasmid fractions in sample set A is 3.97-36.2%. In sample set B, the fraction of ARGs range between 9-30% and the fraction of plasmids range between 60.9-87.7%. Subsequent dilutions yield lesser amounts of post-enrichment DNA (figure 6B). The fraction of plasmids/ARGs is high throughout all samples. The highest representation of genetic diversity is found in sample B, using 10ng of starting material (72% plasmid coverage, 64% ARG coverage).

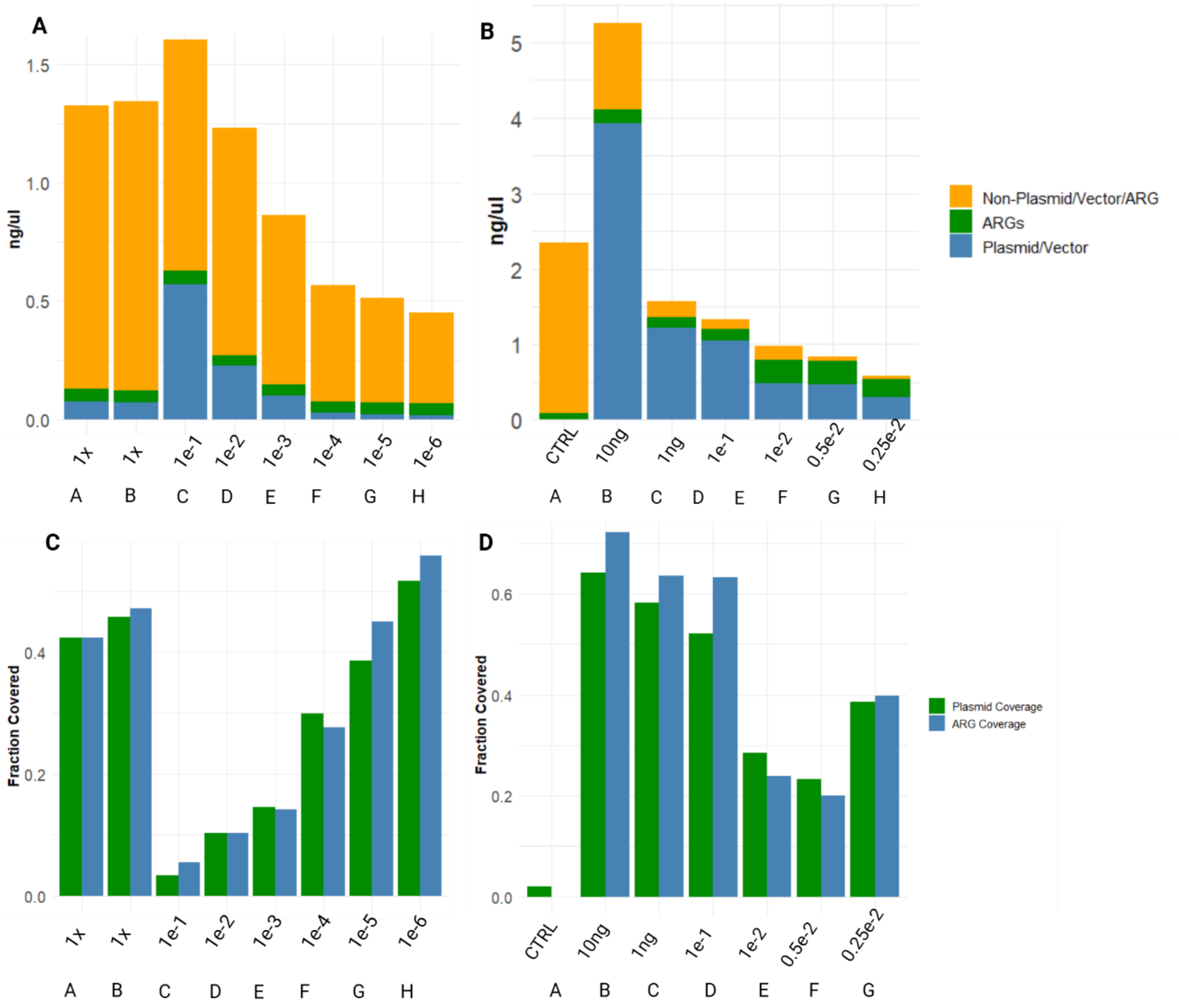

Figure 6. A-B) Abundance of plasmid and non-plasmid throughout the dilution series' (A): Sample set A (fecal samples) (B): Sample set B (spike-in samples). C-D) Plasmid coverage throughout the dilution series, normalized to all plasmids discovered throughout all samples per experiment set. The bar size is proportional to the DNA concentration in the samples after enrichment, giving a measurement of total DNA as well as the concentration of plasmid and non-plasmid sequences.

5. Discussion

5.1 Targeted Enrichment

Assessing the microbial abundance of the sample sets enables us to see trends that may be indicative of what initial DNA masses provides the most informative output. This is an important question, as oversaturation of the bait sequences could yield a biased amplification. Microbial patterns arise throughout the different sample sets. In sample set A, the undiluted samples (A-B) have the same microbial signature as the most diluted samples (F-H). The reason for this may be a heterogeneity in the stock fecal sample, leading to disproportional enrichment of certain phyla, in this case Pseudomonadota, which is highly abundant in all samples. The initial overabundance of Pseudomonadota results in those sequences getting proportionally enriched, saturating the bait panel and disabling the baits from enriching other sequences. Another hypothesis regarding the microbial pattern in sample set A, is that the sequences are misclassified. Kraken2 was employed to classify the sequences, using the confidence threshold of 0.1 (10%). Misclassifications (false positives) are surely present, but the high post-enrichment DNA concentration in sample C is not taken into account (figure 8A). In sample set B, the microbial signature stabilizes from sample C onwards (C-H) (figure 4B). This indicates that the bait panel is effectively enriching sequences in the same pattern using 0.25e-2 – 10ng of starting material (figure 4B). The outcome is in line with our hypothesis about panel oversaturation, where an overabundance of initial DNA saturates the panel.

The abundance of plasmid/vector sequences and ARGs was assessed by first assembling reads into on average 7991 contigs per sample composed of, on average 377 nucleotides. The contigs were megaBLASTed, permitting five alignments per query. Many alignments for each query had low E-values, making true hits hard to distinguish from lower probability hits. The ARG/plasmid coverage was assessed by finding the unique plasmids/ARGs in each sample and assessing the representation of unique elements in the samples.

An interesting trend arises in sample set A as the initial DNA mass is reduced. The fraction of unique plasmids/ARGs represented in each sample increases although the fraction of plasmids/ARGs in the samples decrease. Also, post-enrichment DNA concentration is reduced throughout the dilution series, which can be expected when using less initial DNA mass (figure 8A, 8C). The poorest coverage is found in sample C, where the post-enrichment DNA concentration is the highest and where the microbial signature is the most out-of-pattern. Pseudomonadota is drastically enriched in the sample, possibly leading to a high number of enriched targets of low variance. This would explain the low representation of sequence variation and high post-enrichment DNA concentration. The abundance of background sequences is reduced from 91% to 61.8% throughout sample set A, the lowest noise being observed in sample C.

In sample set B, sample B has significantly increased post-enrichment DNA concentration compared to the other samples .The proportion of ARGs/plasmids is slightly reduced, but the DNA concentration is more than doubled compared to the control (sample A). The microbial abundance of sample B differs from the subsequent diltutions (C-H). Sample B hosts significantly more Chordata sequences. The reason for this may be the same as for sample C in sample set A, where a heterogeneity could explain the out-of-pattern enrichment. The prevalence of ARGs is more varied in sample set B compared to sample set A. The highest fraction of ARGs being found in samples with low pre- and post-enrichment DNA concentrations (C-H). The depletion of background sequences is clear in sample set B, where the non-plasmids/ARGs are completely depleted in the final dilution (sample H). The background sequences are reduced from 90% to 10%. The reason for the background depletion being more effective in sample set B may be due to the different sources of the sample sets. Sample set A is a fecal reference sample containing lots of microorganisms' sequences, while sample set B is a cell line (VERO E6) spiked with plasmids. The background of sample set A may be inherently harder to deplete, as ARG-/plasmid-like sequences may be abundant to a higher extent.

5.2 Future efforts

5.2.1 Expanding the bait set

When utilizing oligonucleotides, their thermodynamic properties are important to keep in mind. An essential property in the use-case of targeted enrichment, is the melting temperature of the baits being homogenous. This in turn yields homogenous enrichment of target sequences. The filtering thresholds were set to achieve between 75000 – 100000 baits. However, the filter thresholds can be tweaked to achieve different bait sets from the same database. Future efforts should include the different filters being weighed against each other to optimize the thermodynamic properties of the bait sequences. In order to increase the overall filtering stringency, a larger database should be curated. When curating the current database, we opted for antibiotic resistance *genes*, not resistances brought by mutations as well as plasmid/vector sequences. Future databases will produce baits with higher coverage if these elements are further expanded. Common vector inserts should be considered as well, such as GFP and other genetic elements with empirical synbio applications. It may be valuable to curate a database of virulence factors as well. All microorganisms contain virulence factors, and its definition is not delimited enough to include all of them. Virulence factors are genetic sequences that aid microorganisms by facilitating cell invasion, immune-evasion and a downstream of other mechanisms.

When filtering the bait set, the following bait properties were taken into account:

Gibbs Free Energy

Gibbs free energy (ΔG) predicts chemical *spontaneity*. In this use-case, it translates to how prone the baits are to form secondary structures. If $\Delta G < 0$, the process is spontaneous. The lower the value (the larger the difference in Gibbs free energy between initial and final conditions), the more likely the secondary structure is to form. There is a wide distribution of ΔG among the bait sequences. In order to mitigate the formation of secondary structures, stringent filtering in this step would enable more baits to bind to their target sequences and subsequently get amplified, as they would not be partaking in any secondary structures. The likelihood of sequences partaking in secondary structures can be assessed by their entropy, a measurement of disorganization. High entropy in the context of oligonucleotides corresponds to sequences that are non-repetitive (disordered). Low entropy on the other hand suggests that the sequences are repetitive and organized. Sequence entropy is a measurement that complements ΔG, as it is also a predictor of secondary structure formation. Sequences with low entropy are more prone to form secondary structures than their more complex counterparts. This in turn would lead to probes becoming unavailable for target capture. In this study, no filtering was conducted on the basis of entropy. Instead, entropy was used to inspect the bait panel.

GC-content and melting temperature

Sequences' melting temperatures are proportional to their GC-contents, which in turn is proportional to their annealing temperatures. If bait sequences' GC-contents vary drastically, the resulting baits will anneal to their targets at different temperatures, in turn, leading to biased enrichment. Limiting the GC-content to 35%<GC-content<75%, reduces the variability in melting temperature. The sequence filtering on the basis of GC-content is limited by the number of initial baits. Using a larger database, more stringent filtering could have been conducted. However, more stringent filtering will also yield some information loss, as some biological sequences indeed have high prevalence of guanine and cytosine. When assessing the stringency of future filtering, assessing the database coverage between different filtering steps would provide an assessment of the loss of information throughout the process.

Blacklist alignments

The blacklist contains the human genome (GCF_000001405.26), chloroplast sequences present in NT and the human mitochondrial genome present in NT (gathered using entrez-direct). Filtering against the blacklist is necessary, as the output data otherwise would be flooded by non-plasmid/non-ARG sequences. By adding multiple genomes to the blacklist, it would encompass a larger variety of unwanted sequences. The necessity for this can be seen in sample set B (sample B), where Chordata have been thoroughly enriched.

5.2.2 Improving target coverage metrics

Target coverage is assessed by local alignment on NT, which is bound to produce false positives. To overcome this, future work can emphasize mapping the captured reads onto the database from which the baits are generated. This will provide a measurement of target coverage without relying on local alignments on large databases.

5.3 Study limitations

5.3.1 Experimental setup

The experimental design does not contain any biological or technical repeats, which is motivated by the experimental nature of this study. However, this leaves the results of this study preliminary. Future studies can statistically assess the trends observed here, where different sample sets yield differences in background depletion at different initial DNA masses.

5.3.2 Limitations of microbial classification

Kraken2 was used for classification. Although it is a popular method, it has its flaws, especially when dealing with highly mutagenic organisms. Kraken2 works by identifying overlapping k-mer signatures to its database. When employing default Kraken2 settings, a large fraction of sequences are unclassified (false negatives). In order to mitigate the prevalence of these, the confidence level of Kraken2 is dropped to 0.1. The confidence level is a threshold that dictates the fraction of k-mers required to classify a read to an organism. A confidence threshold of 0.1 means that 10% of k-mers must map to an organism in order for Kraken2 to classify it as such. Increasing the confidence yields a trade-off between the fraction of unclassified reads (false negatives) and false positives.

5.3 Biodefense implications

Applying in-solution hybridization methodologies enables sequence-based enrichment of microorganisms, providing a high-resolution understanding of the sequences of concern in any sample. This study proves that select phyla are enriched when targeting plasmid/ARG sequences, enabling analysis of sequences that would otherwise be drowned in background noise. The utility of bait panels is not limited to biological weaponry, it can also be applied to characterize the spread of diseases in any sample type. The successful employment of in-solution hybridization in the field of biodefense is a big step from the PCR-based and metagenomics approaches currently employed. A limitation when applying in-solution hybridization is that any sequences that are unrepresented by the bait panel, will not be enriched, likely yielding false negatives. Addressing these limitations require further investigation.

6.References

Madsen JM, Darling RG. Future Biologic and Chemical Weapons * *The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the U.S. Army, U.S. Navy, any other U.S. military organization, or any of the places where Dr. Darling works, or any governmental organization. In: Disaster medicine [Internet]. 2006. p. 424–33. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7152170/

G.W Chritstopher, T.J Cieslak, J.A Pavlin, E.M Eitzem Jr. Biological warfare: a historical perspective JAMA, 278 (1997), pp. 412-417

Riedel S. Biological Warfare and Bioterrorism: A Historical review. Baylor University Medical Center Proceedings [Internet]. 2004 Oct 1;17(4):400–6. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1200679/

Venkatesh S, Memish ZA. Bioterrorism—a new challenge for public health. International Journal of Antimicrobial Agents [Internet]. 2003 Feb 1;21(2):200–6. Available from: https://pubmed.ncbi.nlm.nih.gov/12615387/

Pletcher K. Tokyo subway attack of 1995 | Facts, Background, & AUM Shinrikyo [Internet]. Encyclopedia Britannica. 2010. Available from: https://www.britannica.com/event/Tokyo-subway-attack-of-1995

Jernigan DB, Raghunathan PL, Bell BP, Brechner R, Bresnitz EA, Butler JC, et al. Investigation of Bioterrorism-Related Anthrax, United States, 2001: Epidemiologic findings. Emerging Infectious Diseases [Internet]. 2002 Oct 1;8(10):1019–28. Available from: https://pubmed.ncbi.nlm.nih.gov/12396909/

Feakes D. The Biological Weapons Convention. Revue Scientifique Et Technique De L OIE [Internet]. 2017 Aug 1;36(2):621–8. Available from: https://pubmed.ncbi.nlm.nih.gov/30152458/

Chemical Weapons Convention [cited 2025 Jan 8]. OPCW. Available from: https://www.opcw.org/chemical-weapons-convention

Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction (Biological Weapons Convention). Geneva: United Nations; 1972 [cited 2025 Jan 8]. Available from: https://www.un.org/disarmament/biological-weapons/

Brockmann K, Bauer S, Boulanin V. *Bio Plus X: Arms Control and the Convergence of Biology and Emerging Technologies* [Internet]. Stockholm: Stockholm International Peace Research Institute; 2019 [cited 2025 Jan 8]. Available from: https://www.sipri.org/

Padda IS, Tadi P. Botulinum toxin [Internet]. StatPearls - NCBI Bookshelf. 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK557387/

Witmanowski H, Błochowiak K. The whole truth about botulinum toxin – a review. Advances in Dermatology and Allergology [Internet]. 2019 Feb 27;37(6):853–61. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC7874868/

Tegos GP. Biodefense. Virulence [Internet]. 2013 Nov 15;4(8):740–4. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3925707/

Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. *eClinicalMedicine* [Internet]. 2022 Dec;54:101675. Available from: https://doi.org/10.1016/j.eclinm.2022.101675

Penninx BWJH, Benros ME, Klein RS, Vinkers CH. How COVID-19 shaped mental health: from infection to pandemic effects. Nature Medicine [Internet]. 2022 Oct 1;28(10):2027–37. Available from: https://www.nature.com/articles/s41591-022-02028-2#Tab1

World Health Organization. *Mental health and COVID-19: scientific brief* [Internet]. Geneva: World Health Organization; 2022 Mar 2 [cited 2025 Jan 8]. Available from: https://www.who.int/publications-detail-redirect/WHO-2019-nCoV-Sci-Brief-Mental-health-2022.1

Tegos GP. Biodefense. Virulence [Internet]. 2013 Nov 15;4(8):740–4. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3925707/

Yasumoto S, Muranaka T. Foreign DNA detection in genome-edited potatoes by high-throughput sequencing. Scientific Reports [Internet]. 2023 Aug 9;13(1). Available from: https://www.nature.com/articles/s41598-023-38897-x

Tay AP, Didi K, Wickramarachchi A, Bauer DC, Wilson LOW, Maselko M. Synsor: a tool for alignment-free detection of engineered DNA sequences. Frontiers in Bioengineering and Biotechnology [Internet]. 2024 Jul 12;12. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11272466/

Wu S-L, Tan Y-Y, Zhao Y, Fan L-J, Gao Q-K, Gatehouse AMR, Shu Q-Y. CTREP-finder: A web service for quick identification and visualization of clean transgenic and genome-edited plants. *Crop Design* [Internet]. 2022 Apr 9 [cited 2025 Jan 8];1:100003. Available from: https://doi.org/10.1016/j.cropd.2022.03.001.

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu CH, McDermott PF, Tadesse DA, Morales C, Simmons M, Tillman G, Wasilenko J, Folster JP, Klimke W. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. Antimicrob Agents Chemother. 2019 Oct 22;63(11):e00483-19. doi: 10.1128/AAC.00483-19. Erratum in: Antimicrob Agents Chemother. 2020 Mar 24;64(4): PMID: 31427293; PMCID: PMC6811410.

Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728. doi: 10.1038/s41598-021-91456-0. PMID: 34135355; PMCID: PMC8208984.

Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrobial Agents and Chemotherapy [Internet]. 2013 Oct 22;58(1):212–20. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC3910750/#abstract1

Alcock *et al*. 2023. CARD 2023: Expanded Curation, Support for Machine Learning, and Resistome Prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, 51, D690-D699.

Doster, E., Lakin, S. M., Dean, C. J., Wolfe, C., Young, J. G., Boucher, C., Belk K. E., Noyes N. R., Morley P. S. (2019) MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res. doi:10.1093/nar/gkz1010.

Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microbial Genomics [Internet]. 2022 Jan 17;8(1). Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8914360/

The UniVEC database [Internet]. Available from: https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/#Sources

Zymo Research. ZymoBIOMICS™ Fecal Reference with TruMatrix™ Technology: Instruction Manual. Version 1.0.3. Irvine, CA: Zymo Research Corporation; 2021. Available from: [insert URL if applicable]

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology [Internet]. 2017 Oct 16;35(11):1026–8. Available from: https://www.nature.com/articles/nbt.3988

Appendix

Table SI. Plasmids used to spike the VERO E6 cell line. The ratio (v:v) between plasmids are 1:1, the concentration of each plasmid is different. The plasmid volume was diluted to 100 microliters, then diluted when performing TWIST targeted enrichment.

Backbone	**Insert**
psFUI	RUF62
PSFU3	hoc2
PSCA1	G1-2
PCAGGS	VSV-G
pcDNA3.1	RUFG2-1
pET19B	RNF-N
PET151	S-N1
PET101	RUF-N
PCR2.1	empty
unknown	NSP2
PKK289	empty
PNQ705	empty
PSMP6	empty
PETM30	empty

PETM40	empty
PETM41	empty
PETM20	empty
P0451-S2	crAT
PET9D	empty
pD603	AcHe

Table SII. Microbe overlap between fecal samples.

Microorganism	**Hosted by *n* samples**	**Samples**
Actinomycetota	6	10e-2, 10e-3, 10e-1, 10e-5, 10e-6, 10e-4
Chordata	6	10e-1, 10e-2, 10e-3, 10e-4, 10e-5, 10e-6
Kitrinoviricota	6	10e-2, 10e-1, 10e-3, 10e-4, 10e-6, 10e-5
Pseudomonadota	6	10e-1, 10e-2, 10e-3, 10e-4, 10e-5, 10e-6
Myxococcota	5	10e-2, 10e-3, 10e-5, 10e-6, 10e-4
Patescibacteria	4	10e-2, 10e-1, 10e-6, 10e-3
Uroviricota	4	10e-2, 10e-3, 10e-1, 10e-4

Acidobacteriota	3	10e-3, 10e-5, 10e-6
Bdellovibrionota	3	10e-3, 10e-6, 10e-5
Desulfobacterota	3	10e-5, 10e-4, 10e-6
Bacillota_A	2	10e-5, 10e-6
Bacillota_C	2	10e-5, 10e-6
Bacillota_G	2	10e-5, 10e-6
Cyanobacteriota	2	10e-5, 10e-6
Dependentiae	2	10e-4, 10e-6
Desulfobacterota_B	2	10e-5, 10e-6
Elusimicrobiota	2	10e-6, 10e-4
Fusobacteriota	2	10e-4, 10e-5
KSB1	2	10e-3, 10e-2
Planctomycetota	2	10e-4, 10e-6
Thermotogota	2	10e-4, 10e-6
Verrucomicrobiota	2	10e-5, 10e-6
Bacillota	1	10e-3
Bacillota_B	1	10e-6
Bacillota_D	1	10e-6

Bacillota_E	1	10e-6
Bacillota_I	1	10e-5
CALINM01	1	10e-5
Caldisericota	1	10e-4
Chlamydiota	1	10e-5
Chloroflexota	1	10e-6
Cossaviricota	1	10e-3
Deinococcota	1	10e-4
Desulfobacterota_D	1	10e-2
Eremiobacterota	1	10e-5
Euryarchaeota	1	10e-6
Fermentibacterota	1	10e-6
Gemmatimonadota	1	10e-6
Hydrogenedentota	1	10e-6
Margulisbacteria	1	10e-6
Marinisomatota	1	10e-6
Nitrospirota	1	10e-6
Omnitrophota	1	10e-4

Spirochaetota	1	10e-2
UBA9089	1	10e-6
WOR-3	1	10e-6
Zixibacteria	1	10e-6

Table SIII. Microbe overlap in the spike-in samples.

Microorganism	Hosted by *n* samples	Samples
Bacteroidota	7	0.1ng, 1ng, 10ng, 0.05ng, 0.01ng, 0.025ng, 25ng
Kitrinoviricota	7	1ng, 10ng, 0.1ng, 0.05ng, 0.01ng, 0.025ng, 25ng
Actinomycetota	6	0.01ng, 0.1ng, 0.05ng, 0.025ng, 1ng, 10ng
Deinococcota	6	0.05ng, 10ng, 1ng, 0.1ng, 0.01ng, 0.025ng
Negarnaviricota	6	0.05ng, 0.01ng, 1ng, 0.1ng, 10ng, 0.025ng
Pisuviricota	6	0.05ng, 1ng, 10ng, 0.1ng, 0.01ng, 0.025ng
Pseudomonadota	6	0.01ng, 10ng, 0.1ng, 1ng, 0.05ng, 0.025ng

Verrucomicrobiota	6	0.05ng, 0.1ng, 0.025ng, 10ng, 1ng, 0.01ng
Bacillota	5	0.05ng, 0.01ng, 10ng, 0.1ng, 1ng
Bacillota_A	2	0.01ng, 0.05ng
Atribacterota	1	0.05ng
Peploviricota	1	25ng

Table SIV. DNA concentrations after enrichment for the spike-in samples.

Sample	DNA concentration (ng/ul)	avg Fragment length
25 ng control	2.48	309
25 ng enriched	3.6	325
10 ng enriched	5.22	305
1 ng enriched	1.44	365
0.1 ng enriched	1.2	432
0.01ng enriched	0.8	370
0.005 ng enriched	0.61	363
0.0025	0.376	394

Table SV. DNA concentrations after enrichment for the fecal samples.

Sample	DNA concentration (ng/ul)	avg Fragment length
25 ng (1x)	1.34	264
25 ng (1x)	-	249
2.5 ng (1e-1x)	1.58	252
0.25 ng (1e-2x)	1.22	312
0.025 ng (1e-3x)	0.846	-
0.0025 ng (1e-4x)	0.542	300
0.00025 ng (1e-5x)	0.486	-
0.000025 ng (1e-6x)	0.418	249

Table SVI. The table displayed the fractions of reads classified into distinct categories, the reads were classified using Kraken2.

ID	Sample	Human	Bacterial	Viral	Unclassified
A	25 ng control	79%	17%	0%	4%
B	25 ng enriched	43%	40%	13%	3%
C	10 ng enriched	1%	82%	13%	4%
D	1 ng enriched	0.8%	82%	14%	3%
E	0.1 ng enriched	1%	82%	14%	3%
F	0.05 ng enriched	2%	84%	12%	2%
G	0.025 ng enriched	12%	77%	9%	2%t
H	0.01 ng enriched	3%	83%	12%	2%

Table SVII. The distribution of taxonomy in the fecal dilution series The reads were classified using Kraken2.

ID	**Sample**	**Human**	**Bacterial**	**Viral**	**Unclassified**
A	1x	0.07%	99%	0.04%	0.5%
B	1x	0.06%	100%	0.04%	0.3%
C	e-1	5%	93%	0.2%	2%
D	e-2	4%	94%	0.4%	2%
E	e-3	3%	96%	0.5%	1%
F	e-4	0.7%	99%	0.1%	0.5%
G	e-5	0.1%	99%	0.04%	0.4%
H	e-6	0.1%	99%	0.04%	0.6%

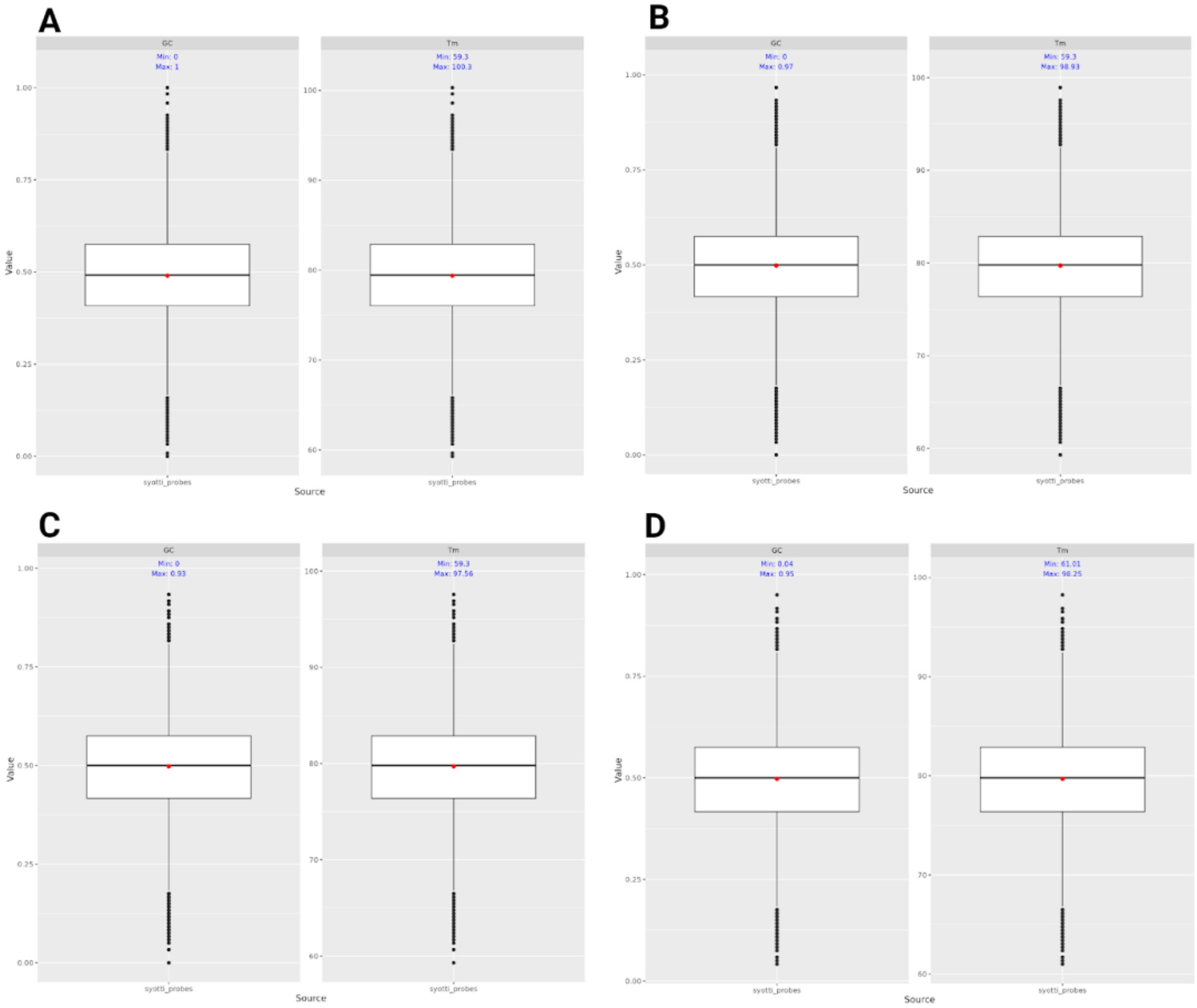

Figure S1: Melting temperatures and GC content for probes generated by Syotti using different settings for maximum allowed mismatches. A) 0 allowed mismatches. B) 10 allowed mismatches. C) 20 allowed mismatches. D) 40 allowed mismatches. Thechoice of mismatches primarily affects the number of probes. The distribution of GC% and melting temperature is roughjly the same in a ll probe sets.