UNIVERSITY OF HELSINKI

Identification of *Staphylococcus* bacteriophage Stab21 toxic gene products using *Escherichia coli* as a host

Master’s programme in Microbiology and Microbial Biotechnology
Master's thesis

Author:
Ellisiv Nyhamar

Supervisors:
Ph.D. Xing Wan
Ph.D. Mikael Skurnik

27.11.2022
Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet/ — Faculty	Masters´s Programme
Faculty of Agriculture and Forestry	Master's programme in Microbiology and Microbial Biotechnology

Tekijä — Författare — Author

Ellisiv Nyhamar

Työn nimi — Arbetets titel — Title

Identification of *Staphylococcus* bacteriophage Stab21 toxic gene products using *Escherichia coli* as a host.

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Master's Thesis	October 2022	51

Tiivistelmä — Referat — Abstract

S. aureus infections are prominent worldwide, and with the rapid increase in antimicrobial resistant variants such as methicillin-resistant MRSA, the need for new treatment alternatives is imminent (Monaco et al., 2017). Lytic bacteriophages are continually evolving new methods for the destruction of bacterial cells while avoiding their defence mechanisms. Screening hypothetical proteins of unknown function (HPUFs) from bacteriophages for toxic activity against bacteria may provide new and potentially life-saving approaches to combat bacterial infections (Liu et al., 2004, Singh et al., 2019).

The Stab21 phage of *Staphylococcus* is a recently described lytic phage with over 85 % of its open reading frames annotated as HPUFs (Oduor et al., 2019). The successful identification of potentially toxic gene products could facilitate the discovery of novel bacterial targets for the development of new antimicrobials. It could also provide treatment options to multi-drug resistant *S. aureus* caused infections where no effective drugs are currently available. To reduce unnecessary screening of phage particle associated yet poorly annotated proteins, total proteins of phage particle were previously identified by LC-MS. Similar studies have previously been performed with *Yersinia* phage fR1-RT and *Klebsiella* phage fHe-Kpn01, where a handful of toxic proteins were discovered (Mohanraj et al., 2019, Spruit et al., 2020). To accelerate the screening process, a next-generation sequencing (NGS) high-throughput screening method was further developed by Kasurinen et al. (2021).

In this study, 96 true HPUFs were selected and screened for their bactericidal activity in *E. coli* using the NGS-based approach. Fourteen potentially bacteriotoxic Stab-21 gene products were identified through toxicity screening in *E. coli*. Of these, three had a particularly low ratio of isolated plasmid after transformation while having a significant number of reads over each joint sequence, indicating their potentially high toxicity. The three most promising candidates were the gene products of *g008*, *g081c* and *g175* of the Stab21 bacteriophage.

Avainsanat — Nyckelord — Keywords

Bacteriophage, antimicrobial resistance, virology, bacteriology, toxicity screening

Säilytyspaikka — Förvaringsställe — Where deposited

HELDA — Digital Repository of the University of Helsinki

Muita tietoja — Övriga uppgifter — Further information

Table of contents

1 Introduction

Antimicrobial resistant bacteria are one of the biggest threats to modern medicine. Since the first antibiotic drug, penicillin, was implemented for treatment of bacterial infections in the 1940s, bacteria have fought back through the development and spread of resistance genes. New antibiotic compounds were later discovered, and new resistant strains followed. However, the introduction of new classes of antibiotics has slowed drastically over the decades, with only two introduced to the market since 1962 (Coates et al., 2011). The need for new antibiotic classes is only increasing as the potential for analogue development from existing classes is depleted. This is poignantly described by the World Health Organization's (WHO) report on the twelve bacterial families for which research and development of new antimicrobial treatments are most urgently needed, classified in the priorities critical, high, and medium (World Health Organization, 2017). Pathogens that exhibit alarming resistance against current antimicrobial treatments are given the acronym of ESCAPE, which includes *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa*, and *Enterobacter* species (Mulani et al., 2019).

One ESCAPE species which also features in the WHO global priority pathogens list is methicillin and vancomycin-resistant *Staphylococcus aureus*, placed in the high priority category (World Health Organization, 2017). *Staphylococcus* species cause disease ranging from minor local infections to lethal systemic infections. *S. aureus* strains are particularly virulent, and infections are prominent worldwide (Miklasińska-Majdanik, 2021). *S. aureus* cells can gain and exchange new resistance genes through horizontal gene transfer where more than one strain is present (Cespedes et al., 2005). This is especially common in healthcare and livestock facilities. In fact, multi-resistant strains were already in 2008 responsible for over half of the deaths caused by healthcare-associated bacterial infections (Watson, 2008). The resulting arms-race between new multi-resistant strains such as methicillin-resistant MRSA and the development of new antimicrobials is an ongoing battle that requires immediate attention (Miklasińska-Majdanik, 2021. Monaco et al., 2017). Two new classes of antibiotic compounds with toxic activity against multi-resistant gram-positive bacteria including *S. aureus* have been discovered in 2015

and 2018, although no drugs from these have yet reached the market (Ling et al., 2015, Hover et al., 2018).

One promising alternative to the failing of conventional antibiotics is the treatment with bacteriophages, also known as phage therapy. Phages are viruses that exclusively infect bacteria. Through various dedicated mechanisms, phages reprogramme the bacterial cellular metabolism to produce their own viral particles (Azam and Tanji, 2019). Lytic phages are the most studied phages for therapeutical applications, as their life cycle rely on the destruction of the host cell to release virions to the environment (Gordillo Altamirano and Barr, 2019). Currently phage therapy is utilized as treatment of antimicrobial-resistant infections and in food and other industry applications with the common purpose of eliminating unwanted bacteria. Phage therapy as a viable clinical treatment option is still in the early stages but has yielded promising results in select cases. Schooley et al. successfully developed a phage-based treatment against a multi-resistant clinical isolate of *Acinetobacter baumannii*, which when administered to the patient cleared a persistent infection. The cocktail of nine lytic *A. baumannii* phages was constructed by in vitro screening of previously harvested phages against the patient isolate (Schooley et al., 2017). A further approach to the development of clinical phage therapy treatments was demonstrated by Dedrick et al. where promising phages from a screening assay were later optimized through genome engineering to treat a resistant *Mycobacterium abscessus* infection (Dedrick et al., 2019).

Elimination of bacterial infections with phages has proved to be a possible alternative where conventional treatment with antibiotics is ineffective. however, the current approach is not without faults. Most notably, phages often have a very narrow host-range, and therefore must be screened and tailored to individual bacterial isolates from a single patient. Standardization of a phage therapy regimen that can be applied to multiple patients without customization is therefore very difficult (Oechslin, 2018). Elimination of administered phages by the patient immune system often necessitates long phage therapy treatment regimens with frequent addition of new phages to maintain a therapeutic level (Schooley et al., 2017). In the case of widespread use of select phages in the treatment of certain infections, resistance to the specific target of the phage can develop just as with antimicrobial compounds (Oechslin, 2018). Phage resistance has already been seen, for example in the multi-resistant strain ST258 of

Klebsiella pneumoniae (Hesse et al., 2020). The use of multi-phage cocktails with multiple cellular targets reduces the incidence of resistance but is still not sufficient to eliminate the issue (Hesse et al., 2020. Oechslin, 2018).

Lytic bacteriophages are continually evolving new methods and optimizing existing systems for the destruction of bacterial cells while avoiding their defense mechanisms. Instead of whole phage particles, individual phage-derived proteins that display bacteriotoxic activity can be isolated and used as antimicrobials (Roach and Donovan, 2015, Schmelcher et al., 2012, Schmelcher and Loessner, 2016). Although phage genomes are vastly diverse, there are several well studied groups of proteins utilized by a wide range of phages. Phage-mediated lysis of bacterial cells depends particularly on holins and endolysins, for lysis of gram-negative bacteria also with the help of spanins (Saier and Reddy, 2015). Virion-associated peptidoglycan hydrolases and polysaccharide depolymerases are also phage-derived proteins studied for their bactericidal properties (Roach and Donovan, 2015).

Phage-inspired antibacterial target discovery is another ascending approach to harnessing the antimicrobial activity of phages (Wan et al., 2021). A large proportion of phage gene products have completely unknown function, as they have never been characterized and have sequences that do not correspond with any proteins of known function. These can be screened on an individual basis to identify specific cellular targets to inhibit virulence factors of the host bacteria. Screening hypothetical proteins of unknown function (HPUFs) from bacteriophages for toxic activity against bacteria may provide new and potentially life-saving approaches to combat bacterial infections (Liu et al., 2004, Singh et al., 2019). Previous screening approaches have identified HPUFs with bacteriotoxic activity from a wide range of phages including mycobacteriophages, *Pseudomonas* phages, and *Staphylococcus* phages (Singh et al., 2019, Van den Bossche et al., 2014, Liu et al., 2004).

The Stab21 lytic *Staphylococcus* phage is a recently discovered phage by Oduor et al (2019). The phage is regarded to be from the *Kayvirus* genus of the Twortvirinae subfamily in the Herelleviridae family (accession number LR215719, Oduor et al., 2019). The double-stranded DNA genome of 153797 base pairs was isolated from a water sample from Shkoder, Albania with the host strain *Staphylococcus xylosus* (Oduor et al., 2019). As a lytic phage, Stab21 possesses genes coding for proteins that directly or indirectly mediate the destruction of its host bacterial cells (Sharma,

2013). An average of 78% nucleotide sequence identity has been found between different human-colonizing *Staphylococcus* species (Takeuchi et al., 2005). It possible that some of the same bacteriotoxic gene products active on the original host strain *S. xylosus* may be active against more clinically relevant Staphylococcus species such as *S. aureus*.

Previously, identification of antibacterial activity from novel phage gene products has been performed through various screening methods. Firstly, plating assays where host bacteria are cloned with individual phage genes to discover gene products with growth-inhibiting properties have yielded promising leads. Liu et al. screened predicted ORFs from genomes of 26 *Staphylococcus aureus* phages in *S. aureus* cells. They identified 31 novel families of growth-inhibiting peptides. Following their identification, cellular targets of several novel polypeptides were determined and the interaction between open reading frame 104 of phage 77 and the putative helicase loader DnaI of *S. aureus* was presented as a promising lead towards a new mechanism of bacteriotoxic activity for future antibiotic compound development (2004). Spruit et al. used a similar initial screening assay to screen 22 HPUF-encoding genes from Phage fHe-Kpn01 of *Klebsiella pneumoniae* in *Escherichia coli* to identify the products of *g10*, *g22*, and *g38* as bacteriotoxic (2020). Unfortunately, the method of cloning with single genes followed by CFU-based assessment of toxicity from plating of individual transformations is inefficient and time consuming, limiting the number of gene products that can be screened.

An alternative is creating phage-gene libraries where small fragments of the phage genome are cloned to a vector and transformed to bacterial cells. Singh et al. applied this method to genomes of seven mycobacteriophages where the library was screened against *Mycobacterium smegmatis* cells. Gp49 of the Che12 phage and Gp34 of the D29 phage were identified as bacteriotoxic from the clones causing growth defects in the bacteria (2019). Shibayama and Dabbs screened a library of phage YF1 gene fragments against *Rhodococcus equi* cells and identified ten fragments with bacteriotoxic activity (2011). This method has the potential for high-throughput screening of genes as no prior selection or production of genes to study is required. Library fragments can be also pooled before transformation, saving time and resources. However, transformants still need to be processed individually for induction of transcription and therefore the number of fragments that can be

screened is limited. Fragments are also not guaranteed to contain complete or functional genes and may contain multiple and unknown (hypothetical) genes, complicating the process (Singh et al., 2019, Shibayama and Dabbs, 2011).

The NGS toxicity screening approach combines the advantages of both individual gene plating assays and genomic libraries (Kasurinen 2021). First, by eliminating the structural proteins which are poorly annotated with LC-MS, only the true hypothetical proteins of unknown function are chosen for screening. This significantly reduces unspecific clones. Second, to reduce the number of necessary electroporation steps to bacteria, the gene constructs are pooled before transformation. Third, bacterial transformants are also pooled and sequenced as a single sample. Through 3 levels of dimension ascension, the workload is dramatically reduced per phage genome, which makes screening of multiple genomes possible in a shorter time. Reads analysis through batch jobs on a supercomputer make the processing of sequencing reads simple and quick, where the necessary data can be extracted in a day.

Among all 176 hypothetical genes from the Stab21 genome, 96 genes encoding hypothetical proteins do not match any previously studied proteins (Appendix 1). Out of these 96 gene products, there is a potential for discovery of bactericidal proteins which can alter the bacterial pathways in an unprecedented manner. Screening their toxicity with a high-throughput method efficiently will generate leads to gene products of particular interest that display potentially bacteriotoxic activity. The identified leads can be further studied for their bacterial targets and mechanism of toxicity. In this study, 14 potentially toxic gene products of Stab21 HPUFs were identified. Of these, Gp008, Gp081c and Gp175 yielded the strongest indication of antimicrobial activity in *E. coli*. In the future, confirmation of toxicity of these 14 candidates using tightly controlled expression in growth curve assays should be performed.

2 Materials and Methods

2.1 Bacterial strain, plasmid and phage DNA

Commercial electrocompetent *Escherichia coli* ElectroMAX™ DH10B was used for transformation (catalog number 18290015, Thermo Fisher Scientific, USA). *E. coli* DH10B and derivatives were grown in Luria-Bertani (LB; 10 g/L Bacto™ Tryptone, 5 g/L Bacto™ Yeast Extract, 10 g/L NaCl) agar or broth or super optimal broth with catabolite repression (SOC) medium (Thermo Fisher Scientific, USA) at 37°C or 35°C with shaking at 200 rpm as stated. LB supplemented with 100µg/mL Ampicillin (Amp100) was used to maintain the plasmids.

The pCU1LK shuttle vector was previously constructed from pCU1 with the insertion of a 45 bp linker fragment in the multiple cloning site at the KpnI - PstI site (Figure 1) (Augustin et al., 1992, unpublished result). pCU1LK was used as the cloning vector in the screening assay in this project. The plasmid contains both *ampR* and *cat* of *E. coli* yielding ampicillin and chloramphenicol resistance respectively, but only ampicillin was used for selection.

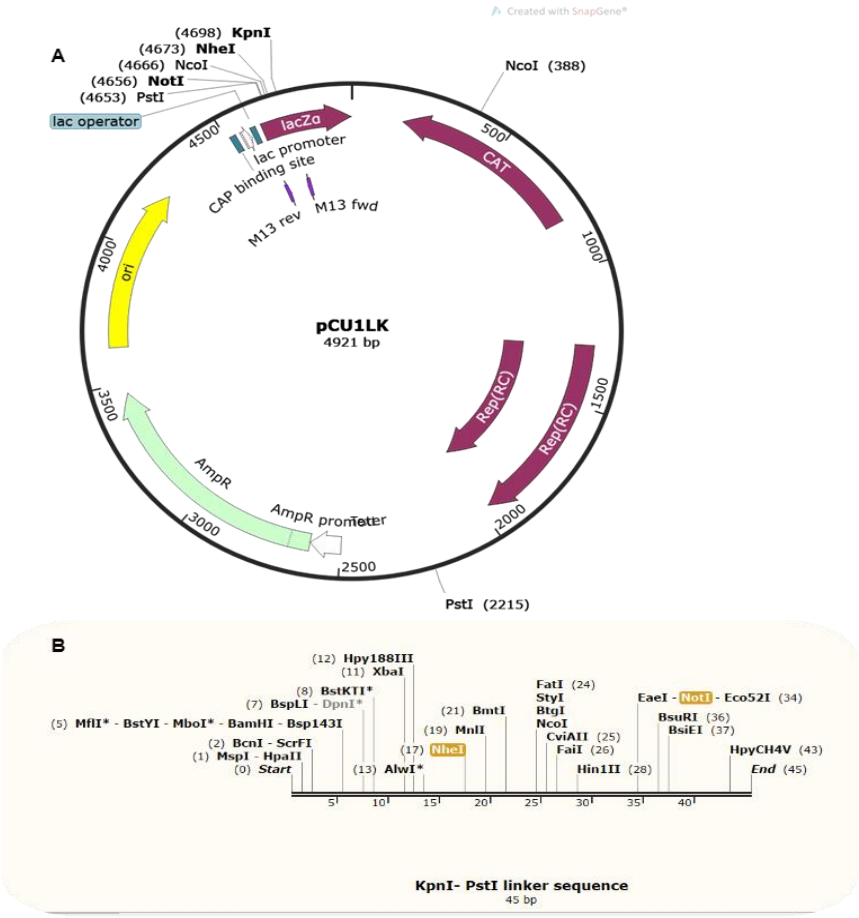


Figure 1. Map of the constructed plasmid pCU1LK (A) where a PstI-KpnI linker fragment (B) has been added to the multiple cloning site of the pCU1 plasmid. The HPUF-encoding genes were cloned in the linker region. Figure created in SnapGene (GSI Biotech).

Out of 176 hypothetical genes, 96 genes encoding hypothetical proteins of unknown function (HPUFs) were selected from the Stab-21 genome using similar approach as previously described by Mohanraj et al. (2019). These HPUFs were previously amplified by PCR and double digested with FastDigest™ restriction enzymes NotI and NehI or KpnI (Thermo Fisher Scientific, USA) (Appendix 1, unpublished result).

2.2 DNA manipulation

E. coli DH10B pCU1LK+ strain was streaked on an LB Amp100 plate to obtain single colonies. One colony was picked at random and inoculated in 200mL LB Amp100 broth. Cells from the overnight culture were collected by centrifugation, pCU1LK vector plasmid was isolated from the cells using NucleoBond Xtra Midi kit according to manufacturer's protocol (MACHEREY-NAGEL, Germany) and was eluted in 200µL Baxter Sterile Water (Baxter Corporation, USA).

pCU1LK vector was linearised using double digestions with restriction enzymes NotI and NheI or KpnI FastDigest™ enzymes (Thermo Fisher Scientific, USA) depending on the insertion fragments (Appendix 1). Both enzyme combinations were incubated in universal FastDigest™ Buffer (Thermo Fisher Scientific, USA). Reaction conditions were 1x reaction buffer, 0.05U/µL of each enzyme, 50 ng/µL DNA and the remaining volume with sterile MilliQ-filtered water. All linearization reactions were incubated at 37°C for one hour before heat inactivation of the enzymes at 80°C for 20 minutes in a T100™ or iCycler Thermal Cycler (Bio-Rad Laboratories, Inc., USA).

Linear pCU1LK vector was dephosphorylated with FastAP™ Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific, USA). The enzyme was added in 0.04U/µL concentration directly to the reaction mixture after linearization and heat-inactivation of the restriction enzymes. The dephosphorylation reaction was incubated at 37°C for 30 minutes followed by 15 minutes heat inactivation at 65°C.

Linear and dephosphorylated pCU1LK vector was ligated with HPUF-encoding gene fragments through sticky-end ligation. T4 DNA ligase and associated buffers (Thermo Fisher Scientific and New England Biolabs® Inc., USA) and sterile MilliQ-filtered water was used in all reactions. To ensure the correct ligation, a 1:3 vector to insert ratio was used, and the total DNA concentration was adjusted to 10 ng/µL. The ligation reaction was incubated at room temperature overnight (15 hours) before heat inactivation at 65°C for 10 minutes.

NucleoSpin Gel and PCR Clean-up kit (REF 740609.50, MACHEREY-NAGEL, Germany) was used to purify and concentrate DNA after enzymatic reactions.

Ligation mixtures and transformant colonies were tested for the correct insertion in PCR systems using vector-embedded primers Puc19-F (`GTCGTGCCAGCTGCAGATCTGAATCGGCCAACGC`) and Puc19-R (`TTCAGCAGAGCTCAGATACCAAAT`) flanking the insertion site. Final concentrations of 1x DreamTaq buffer, 0.05 U/µL DreamTaq DNA Polymerase (Thermo Fisher Scientific, USA), 0.2 mM dNTP Mix, 2 µM of primers, 1 µL of purified ligation mix or cell material from a single colony and sterile MilliQ-filtered water to 50µL volume were used in each reaction.

All PCRs were run in a T100™ or iCycler Thermal Cycler (Bio-Rad Laboratories, Inc., USA) with the program 98°C 30s, (98°C 7s, Tm 20s, 72°C 40s) x34, 72°C 5min, 4°C∞.

For analysis of ligations and colonies with the Puc19 primers an annealing temperature of 60°C was used.

2.3 Analytical methods

2.3.1 DNA analysis

The PCRs and enzymatic restrictions were visualized in 1.0-2.5% agarose gels with 1x TAE as running buffer. GellyPhorLE (Euroclone, Italy) agarose was dissolved in 1x TAE buffer with 0.05% (v/v) Midori Green Advance (NIPPON Genetics Europe, Germany). Final concentration of 1x loading buffer was mixed with the samples and GeneRuler 1 kb DNA Ladder (Thermo Fisher Scientific, USA) was used. The electrophoresis was conducted with a constant potential of 200V and current at maximum 400 mA for approximately 30-50 minutes for a good separation of the DNA fragments. Agarose gels were visualised under UV in a Gel Doc™ XR+ imager with Image Lab™ software (Bio-Rad Laboratories, Inc., USA).

A NanoDrop1000 instrument was used to assess the DNA concentration and the purity of the DNA fragments during the assembly of the gene-vector ligations (Thermo Fisher Scientific, USA). A Qubit instrument was used to measure DNA concentration of ligation mixture pools and transformant plasmid pools more precisely prior to NGS sequencing. Measurements were taken according to the manufacturers' instructions (Thermo Fisher Scientific, USA).

2.3.2 Preparation for Next Generation Sequencing (NGS)

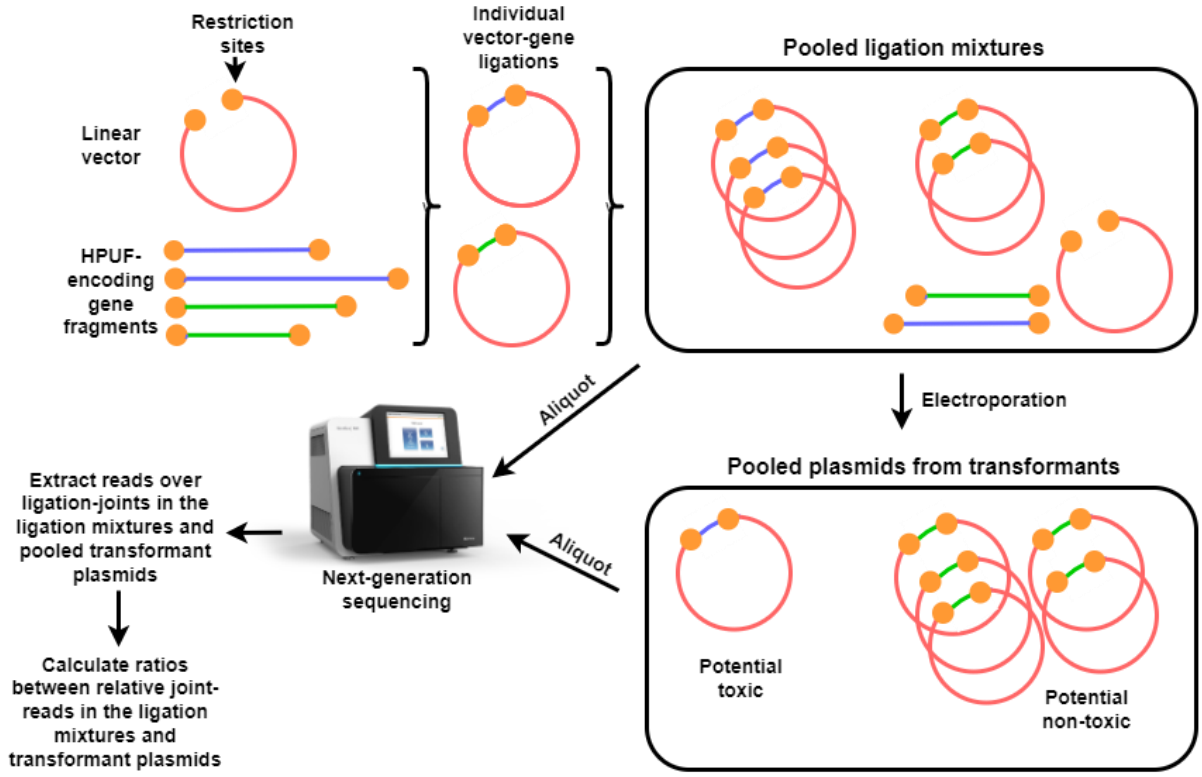


Figure 2. Visual summary of the NGS-based screening method. Adapted from Kasurinen, 2020.

The NGS-screening assay was performed as presented in Figure 2. Every 16 ligation mixtures of HPUF gene and vector pCU1LK were pooled before concentration by kit, and an elution volume of 20µL in Baxter Sterile Water (Baxter Corporation, USA) was used per pool.

One µL (ca. 200ng) of each ligation pool was added to 50µL electrocompetent *E. coli* DH10B cells. Electroporation was performed with a Gene Pulser™ apparatus with the parameters 200Ω resistance, 25mF capacitance and 2.5kV voltage (Bio-Rad Laboratories, Hercules, CA, USA). The cells were recovered in 1mL SOC medium at 35°C for one hour with 15 rpm vertical rotation. Aliquots of 50µL recovered cells were spread on 20 LB Amp100 plates for each pool.

Plasmids from transformation reactions were produced by harvesting all colonies from the transformation plates. 1mL SOC broth was added to each plate and the cells were resuspended with a triangle push rod before transferring to a collection tube. The harvested cell suspension was diluted with SOC with 100µg/mL ampicillin to a total volume of 100 mL and incubated in a 500mL conical flask at 37°C with 220 rpm

shaking for three hours. Plasmids were extracted, purified, and precipitated with NucleoBond™ Xtra Midi kit and NucleoBond™ Finalizers (MACHEREY-NAGEL, Germany) according to manufacturer's instruction. Plasmid pools were eluted in 200µL TRIS/HCl pH=8.5 elution buffer.

DNA samples from both the ligation pool and the plasmid pool were sequenced with the 150 bp paired-end protocol in the Illumina HiSeq platform at NovoGene (UK) as described by Kasurinen et al. (2021) (Figure 2). The 96 chosen genes were grouped in 6 pools of 16, based on previous experience by Kasurinen et al. (2021) where pools of nine to 23 HPUF-encoding genes were used.

2.3.3 *In-silico* analysis

The results from NGS were analysed to reveal the difference between ligation pool and plasmid pool (Figure 2). Ligation-joint sequences were used for screening of the NGS reads for genes that were correctly ligated to the vector (Figure 3). Lists of 28-nucleotide sequences and their reverse complements were generated for each pool by manually compiling both the 3' and 5' gene fragment ends and an equal number of nucleotides from the linear pCU1LK 3' and 5' ends on both sides of the restriction sites (Appendix 2). Thus, a total of four sequences from each gene were listed, as visualized in Figure 3. The sequence hits for each joint were compiled in both the pre- and post-transformation plasmid samples. The reads of each sample were searched for sequences exactly matching the predicted vector-gene joints using the script and workflow developed by Kasurinen et al. (Table S6, Kasurinen et al., 2021).

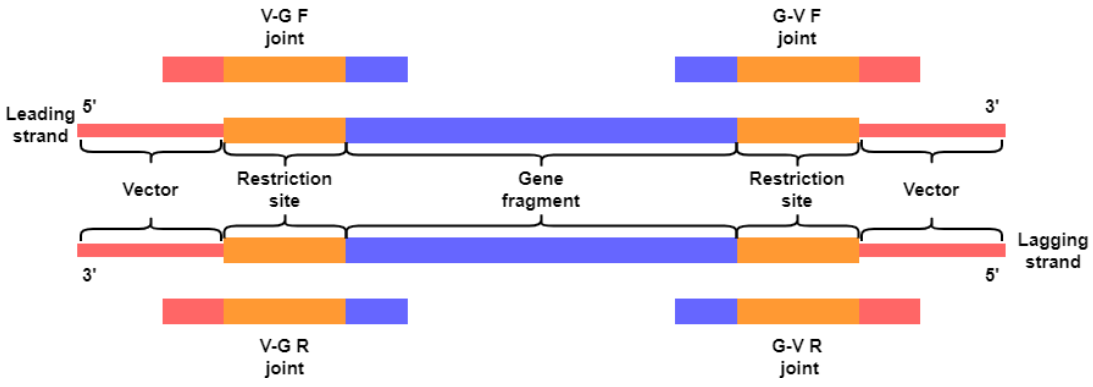


Figure 3. Illustration of the four ligation-joint sequences used in the determination of sequence read coverage for each of the screened HPUFs. Adapted from Kasurinen et al., 2021.

The *in-silico* analysis of raw reads obtained from NGS was conducted on Puhti supercomputer (CSC, Finland) through a windows software PuTTY 0.76 (64-bit, Simon Tatham, 2022), as described previously (Kasurinen et al., 2021). WinSCP 5.19.2 (Windows Secure Copy, Martin Přikryl, 2021) was used to transfer files from Puhti to local file storage.

The relative number of joint reads of each specific gene within a pool was determined by adding the number of reads for all four joint sequences and dividing by the sum of all joint reads in the pool. Potentially toxic genes were identified by ratio lower than 0.5 between the relative joint-reads of a specific gene in the transformant plasmid pool and in those of the ligation mixture pool, as previously established by Kasurinen et al. (2021). The quality of ligations was assessed through sequence alignments of the NGS-obtained reads to the sequence of the HPUF-encoding genes with flanking restriction sites using Geneious Prime 2022.1 (Biomatters, New Zealand). Correlation data analysis, graphs and calculations and were made in Microsoft Excel 2016. Additional figures were constructed in diagrams.net™ (JGraph Ltd, United Kingdom).

3 Results

3.1 Screening of Stab-21 HPUF genes' potential toxicity

To determine the bactericidal activities of the 96 selected HPUFs from the Stab21 phage genome, an NGS-based approach was used which allowed toxicity screening of a high number of gene products in a limited number of assays (Figure 2). In this study, the ligation mixtures were divided and pooled in 6 groups of 16 genes each. Each ligation pool was transferred to *E. coli* through electroporation. Plasmid pools were isolated from the survival transformants of each batch. The ligation mixture pools and plasmid pools were sequenced in NGS to identify the joint-sequences. NGS-reads containing the expected ligation-joint sequences from ligation mixture pools and plasmid pools isolated from transformants were compared (Figure 2). As described in a previous study (Kasurinen 2021), genes encoding a toxic product will not form a plasmid in the transformants, and would therefore have no or little reads over the joint sequence. Transformants carrying a non-toxic gene could conversely yield a high ratio calculated as relative joint-sequence reads of individual genes from plasmid pools over those from the respective ligation mixture. Hence, if the ratio is lower than 1, the gene product can be considered bactericidal. Here, ratios between 0.5 and 1.0 were regarded as potentially mildly toxic, and the cut-off of under 0.5 was used to designate ratios indicating potentially bacteriotoxic activity.

The illumina sequencing yielded between 3.1 and 8.0 million reads from each sample. The number of reads containing each of the predicted ligation-joint sequences were then identified and compiled by a script (Table S6, Kasurinen et al., 2021). The sum of four ligation-joint sequences from each gene were calculated and used as their total read coverage (Table 1, Appendix 2 and 3). The relative number of joint-sequence reads was calculated for all genes by dividing total read coverage of a single gene by the total number of joint-sequence reads for all genes in the pool and expressed as a percentage (Table 1, Appendix 3).

Individual gene read counts varied between ~450 (1.3%) and ~8000 (16%) in the ligation mixtures, and between ~150 (0.03%) and ~1 x 10^5 (18%) in the transformant plasmids (Table 1, Appendix 3). Four genes fell outside this range: *g002*, *g018*, *g062c*, and *g085*. The first two had no detected ligation-joint reads in the plasmid mixture. The read coverage of *g062c* and *g085* in the ligation mixture was exclusively

from the 5' end of the genes, with no ligation joints of the 3' end detected (Table 3). It is probable that an error occurred during enzymatic digestion of these DNA fragments in the preparation of the *g002*, *g018*, *g062c*, and *g085* samples, that likely caused incomplete or wrongly ligated plasmids. The marginal numbers (0-23) of the coverage over the ligation joints detected in the transformant plasmid samples from these four genes are possibly sequencing noise, and not reliable to use in the determination of toxicity. These 4 outliers were therefore excluded from further analysis.

An analysis of the correlation between the relative read coverages for each gene in the ligation mixtures and plasmids isolated from transformants was performed to investigate whether the relative read calculations were skewed by disproportionate representation of any single genes in either sample type. The correlation between the relative reads in the ligation mixtures and plasmids when including every gene was only 0.15.

Table 1. Total sequence reads and relative quantity (%) of reads in ligation mixtures and transformant plasmids for all potentially toxic, potentially mildly toxic, and inconclusive gene products. Relative reads are calculated for each gene from the total number of sequence-reads within each sample from respective pools. Ratios are calculated from relative plasmid reads divided by relative ligation reads. Outlier reads with inconclusive data are in grey. Data from presumed non-toxic genes can be found in Appendix 3.

Gene	Total ligation reads	Relative ligation reads (%)	Total plasmid reads	Relative plasmid reads (%)	Ratio
g002	0	0.0	23	0.0	-
g005	4293	10.3	47398	7.3	0.7
g006	3574	8.6	53519	8.3	1.0
g008	5146	12.4	4225	0.7	0.1
g012	4230	10.2	60401	9.3	0.9
g017	3506	8.4	43185	6.7	0.8
g018	0	0.0	0	0.0	-
g021	2309	5.6	24669	3.8	0.7
g024	3073	7.4	20242	3.1	0.4
g029c	1820	6.4	19606	4.1	0.6
g030c	1934	6.8	20833	4.4	0.6
g031c	1844	6.5	15614	3.3	0.5
g034c	2038	7.2	27444	5.8	0.8
g042c	3107	11.0	36388	7.7	0.7
g044c	1909	6.7	21381	4.5	0.7

g046c	2066	7.3	19514	4.1	0.6
g053c	2352	8.3	22867	4.8	0.6
g061c	3121	14.0	48959	9.5	0.7
g062c	1779	8.0	7	0.0	0.0
g075c	2709	12.2	27433	5.3	0.5
g078c	481	2.2	10109	2.0	0.9
g079c	762	3.4	12744	2.5	0.7
g081c	1324	6.0	152	0.0	0.0
g083c	1334	6.0	20070	3.9	0.7
g085	359	1.6	2	0.0	0.0
g092	2818	12.7	50908	9.9	0.8
g107	1185	3.0	19181	2.3	0.8
g109	3423	8.5	43901	5.2	0.6
g131	3804	9.5	77274	9.1	1.0
g135	4479	11.2	80667	9.5	0.9
g136	2454	6.1	39885	4.7	0.8
g141	3551	8.8	68784	8.1	0.9
g150	4254	10.6	65226	7.7	0.7
g156	4176	10.4	19939	2.4	0.2
g159	3467	8.6	22327	2.6	0.3
g172	8264	15.5	15628	2.4	0.2
g175	7165	13.4	5951	0.9	0.1
g177	2476	4.6	19282	3.0	0.6
g187	4136	7.7	8795	1.4	0.2
g190	3978	7.4	33346	5.2	0.7
g202	792	3.0	12815	2.0	0.7
g204	2063	7.8	35729	5.7	0.7
g206	2342	8.8	50700	8.1	0.9
g209	1545	5.8	14473	2.3	0.4
g211	1799	6.8	30068	4.8	0.7
g212	1788	6.7	7945	1.3	0.2
g213	2969	11.2	27335	4.3	0.4
g215	1164	4.4	8494	1.3	0.3
g216	1997	7.5	7932	1.3	0.2

3.2 Joint-sequence ratios of the potentially toxic genes

The toxicity-screening results of 96 HPUFs from the Stab21 bacteriophage is summarized in Table 2 and Figure 4. The assay identified 14 potentially toxic gene products with ratios under 0.5 between relative ligation-joint reads from the

transformant plasmids and those of the ligation mixture. The genes *g081c*, *g008* and *g175* resulted in the lowest three ratios. An additional 31 HPUF-encoded gene products were identified as potentially mildly toxic with ratios between 1.0 and 0.5. The remaining 47 genes with ratios above 1.0 are considered non-toxic. Previously described genes *g002, g018, g062c,* and *g085* with insufficient data due to errors in sample preparation were deemed inconclusive. The distribution of potentially toxic, mildly toxic, and non-toxic genes was relatively even between the six pools. Only pool number two had no potentially toxic genes identified, and pool six had the most with five.

The ratios of relative reads for all the screened HPUFs varied between 0.005 for *g081c* and 8.51 for *g196*. In the potentially toxic genes, relative read quantities maximally decreased by 99.5% (*g081c*) from 6.0% in the ligation mixture sample to just 0.005% in the transformant plasmid (Table 1). The relative read quantities decreased by between 49.4% (*g031c*) and 3.8% (*g006*) in the potentially mildly toxic genes. Exact ratios and the relative vector-gene joint reads in both ligation mixtures and transformant plasmids for all genes are listed in Appendix 3.

Table 2. Visual summary of the NGS-based screening results. The ratios of relative ligation-joint reads from samples of plasmids isolated from surviving transformants to relative ligation-joint reads from ligation mixtures for all screened hypothetical genes are grouped by presumed toxicity.

Pool 1	Pool 2	Pool 3	Pool 4	Pool 5	Pool 6	
g002	g027c	g061c	g103	g172	g195	**>1.0 (non-toxic)**
g003	g028c	g062c	g107	g173	g196	**1.0-0.5 (potential mildly toxic)**
g005	g029c	g065c	g108	g175	g198	**<0.5 (potential toxic)**
g006	g030c	g069c	g109	g176	g199	**Inconclusive**
g008	g031c	g072c	g131	g177	g201	
g009	g033c	g075c	g135	g179	g202	
g010c	g034c	g078c	g136	g181	g204	
g012	g039c	g079c	g141	g182	g206	
g013	g040c	g080c	g146	g183	g208	
g017	g042c	g081c	g150	g185	g209	
g018	g043c	g083c	g155	g187	g210	
g020	g044c	g085	g156	g188	g211	
g021	g045c	g086	g159	g190	g212	
g022	g046c	g089	g166	g191	g213	
g024	g053c	g092	g169	g192	g215	
g026	g056c	g093	g171	g194	g216	

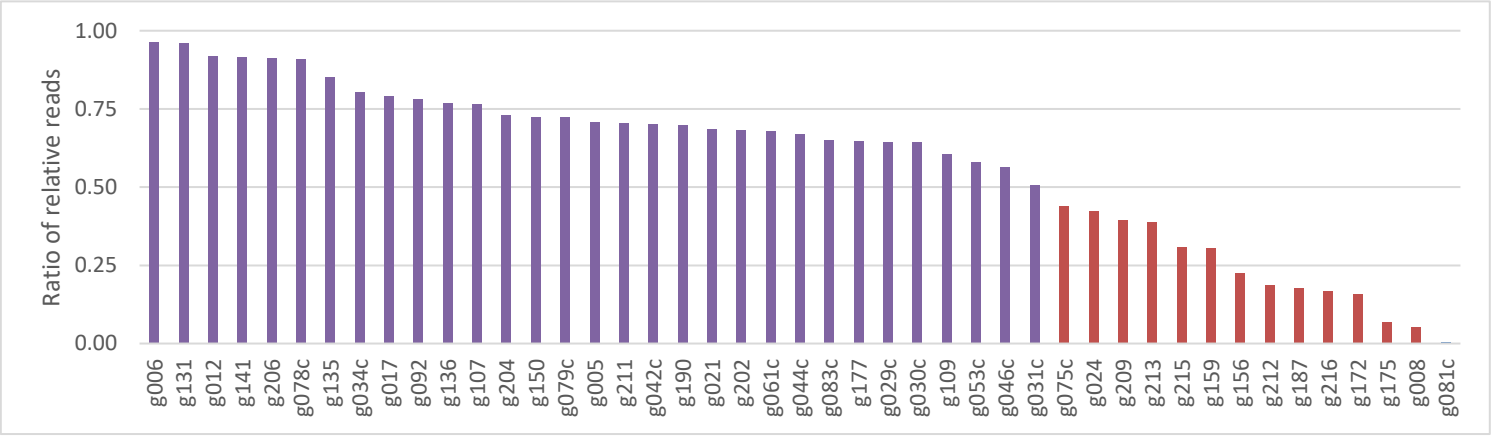


Figure 4. Results from the toxicity screening of Stab21 phage HPUFs. Ratios of relative reads for all 14 potentially toxic (red bars) or 31 mildly toxic (purple bars) genes screened are ordered from highest to lowest. Data for the putative non-toxic genes can be found in Appendix 4.

In the ligation reactions between the linearized vector and gene fragment inserts, the molar ratios between vector and inserts and total DNA concentration was kept constant. Therefore, in theory, the sum of joint reads for a single gene should therefore be 1/16 or 6.25% of the total joint-reads in the pool. In reality, however, there was large variation in the relative number of successful ligations to the vector, especially in cases of the 14 potentially toxic genes (Figure 5). Of the 14 potentially toxic gene products, the majority had above theoretical average of joint reads in the ligation mixture. Although, *g081c*, *g209*, and *g215* had less successful ligation events than the theoretical average, they still produced sufficient ligation-joint reads to be detected for a reliable analysis, with relative joint-read coverages of 6.0%, 5.8%, and 4.39%, respectively.

Similarly, the theoretical average of relative joint-sequence reads of each plasmid from the transformants pool is also 6.25 % in each group, if all the plasmids did not adversely affect the cell growth. A below average percentage of relative joint-sequence reads is expected if background expression is present and potentially toxic gene products produced are killing or inhibiting the growth of transformants before they can produce numerous plasmids. All 14 potentially toxic genes had below average number of joint reads in the plasmids isolated from transformants. This is in contrast to the remaining cases, where 48 of the 78 potentially mildly toxic and non-toxic genes had above-average joint-sequence reads in the plasmid samples.

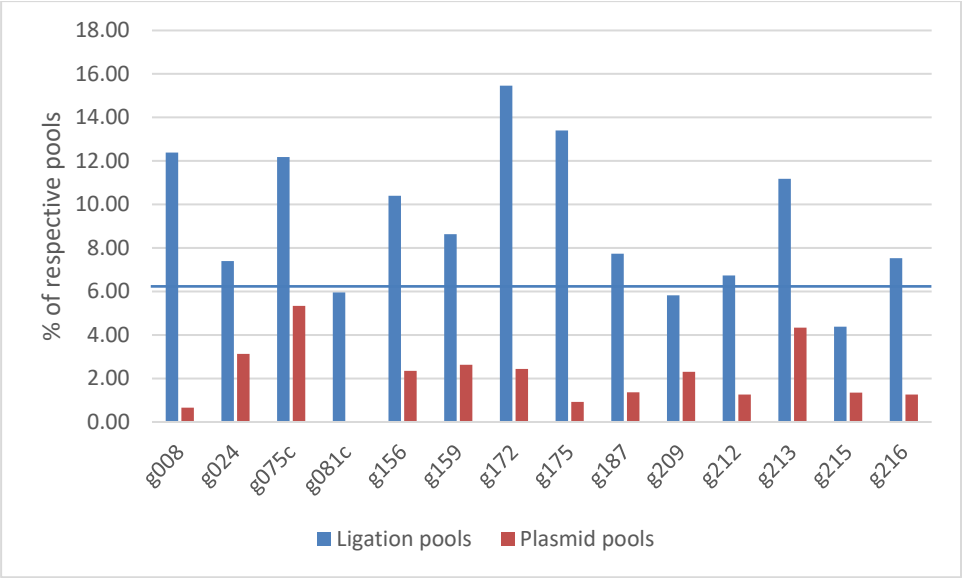


Figure 5. Relative joint-reads of the potentially toxic genes. The presence of the genes identified as potentially toxic in ligation mixtures (blue bars) and plasmids from surviving transformants (red bars) are compared for each gene. The dark blue line at 6.25% (1/16) represents the theoretical value of a single gene if all ligations in a single pool were equally successful and equally toxic. Data from the remaining HPUF-encoding genes can be found in Appendix 5.

To assess the reliability of the NGS-based approach in the screening of bactericidal HPUFs, a closer analysis of the number of reads for individual joint sequences was performed. The standard deviation (SD) was calculated from both the 4 joint reads as an entity and the forward and reverse reads over the joint sequences separately (Table 3).

In general, the reads of the 4 joint sequences had SD values over 500 of the HPUFs in ligation mixture pools, whereas the read counts from plasmid samples were generally close to the mean with only four genes exceeding an SD of 500. For example, *g187* has a standard variation of 1104 in the reads from the ligation mixture, but only 410 from the transformant plasmids. An exception to this trend is *g213*, for which the SD is 654 from the ligation mixture and 1051 from the plasmids. This is mainly due to its high number of reads from the plasmids, as the relative standard deviation expressed as the coefficient of variation percentage (%CV) from the ligation mixture (0.88%) and plasmids (0.15%) still reveal a much higher variation in the ligation mixture reads.

From ligation mixture pools, the read counts of individual gene varied substantially, nevertheless the read coverages remained more or less consistent in either direction. For example, although *g159* in the ligation mixture had a total SD of 970, it was only

95 in the forward reads and 20 in the reverse reads, showing the repeatability of the sequencing events regarding the sequencing direction. An exception was *g008*, which reads varied a lot in either of the sequencing directions but were surprisingly consistent over the joint-sequences. From the plasmid reads, the differences between joint read in either direction were generally smaller than that from the ligation mixture, with SD varying from 22 to 633, excluding the lowest and highest variants *g081c* (3) and *g213* (1072). As the total number of reads in the plasmid samples were generally higher than in the samples from the ligation mixtures, a similar SD means the relative standard deviation was consistently lower in the plasmid samples.

Another variation in the joint reads is the tendency for higher counts of forward reads than the reverse reads in most of the ligation mixtures, especially over the joint sequences covering the 5' end of the genes (Table 3). Examples of this phenomenon are *g156* and *g024* with 4092 and 3001 total reads in the forward direction and only 84 and 72 total reverse reads respectively. Other genes such as *g175* and *g172* have a much more even distribution of reads. This is likely partly because reverse reads tend to have substantially lower quality than the forward reads in Illumina paired-end sequencing and automatic quality control in the sequencing pipeline may remove the majority of the reverse reads with low quality (Kwon et al., 2013). There may also be an artefact introduced from the NGS fragment library generation, where the short length of the gene fragments can cause ligation joints from vector to gene to be detected more often than joints from gene to vector (Kasurinen *et al.,* 2021). However, this probably does not explain the variation fully, as it affects the samples very unequally in the ligation mixtures and much less so in the transformant plasmid samples.

Table 3. Number of ligation-joint reads for individual joint sequences of the 18 HPUFs with the lowest ratio between the relative correct joints in the transformant plasmid pool compared to the relative correct joints in the ligation mixture pool. The genes are listed in order of ratio from lowest to highest. Data from the remaining HPUF-encoding genes can be found in appendix 2. V-G: 28-nucleotide joint sequence covering the transition between vector and the 5' end of the gene with 10 nucleotides on each side of the 8-nucleotide restriction site. G-V: 28-nucleotide joint sequence covering the transition between the 3' end of the gene and into the vector with 11 nucleotides on either side of the 6-nucleotide restriction site. F: forward direction, R: reverse direction. Results in grey constitute genes with insufficient data to determine toxicity.

	From ligation mixture							From transformant plasmids						
Gene	V-G F	V-G R	G-V F	G-V R	SD all reads	SD F reads	SD R reads	V-G F	V-G R	G-V F	G-V R	SD all reads	SD F reads	SD R reads
g002	0	0	0	0	0	0	0	0	23	0	0	12	0	12
g018	0	0	0	0	0	0	0	0	0	0	0	0	0	0

g062c	779	1000	0	0	521	390	500	5	0	0	2	2	3	1
g085	320	39	0	0	155	160	20	1	1	0	0	1	1	1
g081c	349	2	928	45	427	290	22	42	35	35	40	4	4	3
g008	342	380	2805	1619	1173	1232	620	1058	948	1014	1205	109	22	129
g175	2655	1501	1505	1504	576	575	2	1577	1506	1287	1581	138	145	38
g172	2563	1953	1876	1872	333	344	41	3951	3861	3654	4162	211	149	151
g216	589	64	1079	265	443	245	101	2226	1729	1937	2040	207	145	156
g187	1979	27	2000	130	1104	11	52	2574	1955	1748	2518	410	413	282
g212	509	62	996	221	410	244	80	2261	1753	1831	2100	236	215	174
g156	1958	6	2134	78	1160	88	36	5361	4467	4467	5644	609	447	589
g159	1799	10	1609	49	970	95	20	5820	5161	5529	5817	312	146	328
g215	216	68	755	125	315	270	29	2012	1869	1857	2756	428	78	444
g213	974	179	1562	254	654	294	38	6394	6264	6269	8408	1051	63	1072
g209	269	4	620	652	308	176	324	3734	2978	3646	4115	473	44	569
g024	1510	9	1491	63	846	10	27	5473	4309	5106	5354	524	184	523
g075c	494	5	1984	226	894	745	111	7644	6296	6379	7114	640	633	409

3.3 Method efficiency

Transformant colonies were screened with PCR to determine the rate of successful ligations between the HPUF-encoding gene fragments and the vector. The primers flanking the cloning sites in the vector should produce a DNA fragment of 318bp when no gene fragment is inserted. PCR fragments with the approximate length of individual gene fragments plus 318bp were considered positive for a successful ligation event, as ligated plasmids should yield PCR products of 489bp to 1071bp (Figure 6). Early attempts at creating the ligation mixture pools resulted in a high rate of empty vector clones, where no insertion fragment was detected. This led to a dilution of the relevant NGS reads, which made interpretation of the ligation-joint sequences impossible. The DNA manipulation steps were then optimized by changing the purification method of the linearized vector and adding a dephosphorylation step before the ligation reaction to prevent self-ligation of the vector. Both steps significantly improved the ligation quality, which was reflected as good quality NGS reads as shown in Table 3 above.

PCR screening of 16 randomly selected transformants from one optimized ligation mixture pool showed that only 4 contained the empty pCU1LK plasmid, 12 clearly positive clones contained a gene fragment in the cloning site (Figure 6). The calculated rate of 75% useful clones for the toxicity screening assay was sufficient to generate usable data from NGS. Clones 1 and 8 displayed multiple bands in the colony screening, which may stem from issues with the PCR system or a

contamination with cells from more than one colony of transformants with different insert genes in the same reaction. Particularly clone 1 with two equally well-defined bands may be an example of the latter (Figure 6).

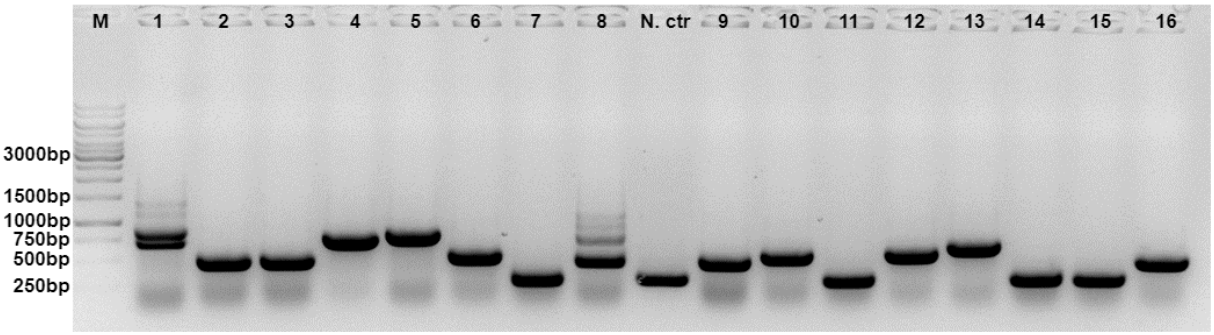


Figure 6. Agarose gel of colony PCR screening to determine rate of successfully ligated gene fragments. Clones 1-16 were randomly selected from transformation plates. The negative control PCR product of the empty pCU1LK vector at 318bp was added to the centre well. All clones with bands larger than the negative control band were considered positive with an insertion of a gene fragment in the cloning site.

A small subset of reads was mapped to the sequences of hypothetical genes *g008*, *g085*, and *g175* from Stab21 to elucidate the reasons for discrepancies in the distribution of sequence reads from the ligation mixes. One million paired reads from respective pools were extracted and mapped to the sequences of *g008*, *g085*, and *g175* from the Stab21 genome (Figure 7, accession number LR215719, Oduor et al., 2019). *g008* was chosen because it had an order of magnitude difference between reads containing the 5' ligation joint and those containing the 3' ligation joint (Table 3). It was found from mapping of the NGS reads that the 5' end of the gene *g008* had an abrupt cut-off at the NotI restriction site in the majority of reads, as the sequence from only a few successful constructs continued over the restriction site to the vector (Figure 7 A). *g175* is included as an example of a ligation with a relatively even distribution of detected reads matching the desired ligation joints (Table 3). From the mapping of reads, it was shown that a large number of reads came from the gene between the restriction sites, but some reads also covered the vector on the flanking regions, indicating successful ligations (Figure 7 B). *g085* was investigated due to the possible failure of the 3' ligation. As shown in Table 3, no reads were detected over the joint sequences over stop codon, and only very few could be detected at the 5' end of the gene. This observation was confirmed through mapping the NGS reads from the ligation mixture to the *g085* sequence, as the reads stopped dramatically at the

restriction sites flanking the gene (Figure 7 C). Full images of the NGS reads mapping with additional genes investigated are in Appendix 6.

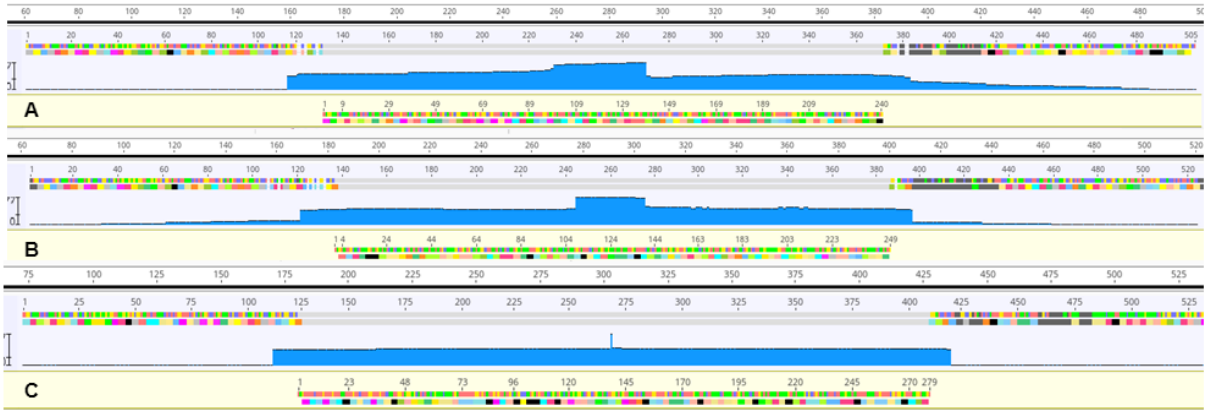


Figure 7. Mapping of NGS reads to sequences of HPUF-encoding genes from the Stab21 phage. A: *g008.* B: *g175.* C: *g085.* Screenshots from Geneious Prime 2022.1. Full versions of the images in Appendix 6.

4 Conclusion and Discussion

The need for new antimicrobial molecules is ever-present and increasing, as bacterial evolution drives forward resistance mechanisms against currently used antibiotics. Strains of common and severe pathogens such as *Staphylococcus aureus*, *Klebsiella pneumoniae* and *Salmonella* species have developed resistance genes against even the latest developed antibiotic treatments available (Prestinaci et al., 2015). Phages provide a promising source of new antimicrobial molecules as the efficiency of traditional antibiotics dwindles. Phage-derived molecules such as endolysins, holins, virion-associated peptidoglycan hydrolases, and polysaccharide depolymerases have all been studied for their bacteriotoxic activity as potential treatments for resistant bacterial infections (Schmelcher et al., 2012, Saier and Reddy, 2015, Roach and Donovan, 2015). However, the study of phage genomes is still in the very early stages with the vast majority of genomes remaining uncharacterized. Of those sequenced, most genes are purely hypothetical and without known function. The staggering variety of phage genomes even within phages of the same host, that can share little to no sequence similarity, provides ample potential for the discovery of new bacteriotoxic gene products that may yield novel antimicrobial mechanisms and targets (Chaitanya, 2019). Gene products with bacteriotoxic activity may provide new leads to the discovery and development of new antibiotic compounds, which could alleviate the impending crisis presented by increasingly antimicrobial-resistant bacterial strains (Wan et al., 2021).

In this study, 96 hypothetical genes of unknown function (HPUFs) from the Stab21 bacteriophage were screened for bacteriotoxic activity against *E. coli*. From the screening, 14 HPUFs were found to be potentially toxic. This is in line with other screening assays of unknown phage genes. Previously, Kasurinen et al. identified 6 potentially toxic genes were from 32 total HPUFs of the vB_EcoM_fHy-Eco03 *Escherichia* phage genome using a similar NGS-based screening assay (2021). Mohanraj et al. identified 8 potentially toxic genes of 94 HPUFs from *Yersinia* phage φR1-RT, and Spruit et al. found 5 in the 22 HPUFs screened from the fHe-Kpn01 *Klebsiella* phage using CFU-based plating assays (2019, 2020). The toxicity screening assay based on next-generation sequencing in this study is a new development which has only been applied to genes of an *E. coli* phage previously (Kasurinen et al., 2021). The experience gained through the herein described assay provides further evidence

that this method is applicable to a wider variety of phages and serves as a valuable starting point in the identification of novel bacteriotoxic peptides.

4.1 NGS-based screening

Using an NGS-based screening approach to screen for phage-encoded toxic proteins is shown to be as reliable as the alternative plating-based toxicity screening method. An earlier direct comparison was performed during development of the NGS method by Kasurinen et al., where it was concluded that the NGS-based assay not only provide similar screening results as the plating-based assays, but also was superior in efficiency, accuracy, and reliability (2021).

An advantage of the NGS-based screening method is the use of ligation-joint sequences as the basis for NGS reads interpretation. As opposed to intra-gene or intra-vector sequences, ligation-joint sequences eliminate the possibility for non-ligated or incorrectly ligated gene fragments and vectors to interfere with the results (Figure 3). This is a direct improvement upon the alternative CFU-based assay, where colonies producing incorrectly ligated, undigested, or self-ligated plasmids constitute false positives (Mohanraj et al., 2019). However, the used of ligation-joints to determine the rate of successful ligations introduces the possibility for false positive results from multiple insertions or ligations that are only successful on one end of the gene. This was seen for example in cases of g062c and g085, where only the 5' end of the gene was ligated correctly to the vector. These one-sided ligated plasmids are not multiplied in transformants, and hence yield a false low ratio between ligation-joint sequence matching reads in the plasmids and those of the ligation mixture. Therefore, this source of error from a disproportion between the reads over the two joint sequences should be identified as outliers and eliminated from the data analysis. Identification of multiple insertions is more difficult. However, as shown in Figure 6, when screened for gene fragment insertion, most colonies carried plasmids with single gene insertion, so the source of error from the wrong clones can be considered minor.

The pools of sixteen genes minimizes the impact of a single gene on the relative abundance of all genes in the pool. The sixteen-gene pools were still small enough to achieve sufficient read coverages from NGS over each joint sequence in both the ligation mixtures and the transformant plasmids. The exceptions were *g002* and

g018 where no correct ligation-joints were found, and *g085* and *g062c* where only one-sided ligation joints were found (Table 3). It is possible that enzymatic restriction of these gene fragments was not completed, hence they could possess incorrect or no restriction site for successful ligation to the vector. As the bioinformatic analysis only used specific search patterns in fuzznuc, these unexpected joint sequences could not be identified. Nevertheless, as these four genes were not from the same pools and each ligation was done individually, their sequencing results did not affect the others. No evidence for significant impact of over- or under-represented genes in the ligation mixture pools was found, as correlation between the relative ligation mixture joint-sequence reads and joint-sequences from the transformant plasmids was only 0.15. In this study, only genes with relative ratios between reads over joint sequences in the plasmid and ligation mixtures under 0.5 were regarded as potentially toxic, although in theory all ratios less than one should indicate the cells having trouble producing the gene products. Using a lower cut-off ratio may exclude some truly toxic gene products with relative ratios between 0.5 and 1, but the rate of false positives is reduced.

The previous application of the NGS-based screening assay included two biological replicates. The variation coefficient between the replicates was consistently lower than that of the comparative CFU-based assay, and stayed under 10% in most cases (Kasurinen, 2020). Therefore, in this study it was decided that the value of increased reliability by including replicates was not higher than the financial and labour costs required. A similar consideration was done on the exclusion of control genes with known toxicity or lack of toxicity. Initial experiments with known genes encoding toxic products in the pCU1LK vector yielded dubious results which required extensive optimization. Considering the scope of a master's thesis, removing control genes from the tests allowed for the full screening of all 96 HPUFs from the Stab21 genome, which was considered more scientifically valuable.

4.2 *E. coli* as screening host

Despite the gene products to be screened originating from a *Staphylococcus* phage, *E. coli* was used due to time constraints, as the protocol for NGS-based screening in *E. coli* was already developed (Kasurinen et al., 2021). *E. coli* as a model organism is well established and can provide insight in cellular mechanisms well beyond the

species itself (Ruiz and Silhavy, 2022). It was also of interest to determine any bacteriotoxic activity in both gram-negative and gram-positive strains. The *E. coli* ElectroMAX™ DH10B strain was chosen due to its high transformation efficiency of over 1 x 10^{10} transformants per μg of DNA. The strain also contains an *end*A1 mutation resulting in increased production and quality of plasmid, which may amplify the bacteriotoxic effect of any toxic gene products present and provide high quality samples for sequencing (Durfee et al., 2008).

The pCU1-based cloning vector in this study was chosen for its compatibility with transformation to both gram-negative *E. coli* and gram-positive *S. aureus* cells (Kim et al., 1994, Uchiyama et al., 2014). In theory, *E. coli* and *S. aureus* share common targets which Stab21 encoded hypothetical proteins could interact with. Hence, using *E. coli* as screening host provides hints on how Stab21 may reprogram the bacterial cells while taking full advantage of the existing biotechnology toolbox designed for the model microorganism. However, if *E. coli* was shown to be entirely immune to any Stab21 HPUFs, the same constructs in this study could be transferred to *S. aureus,* and toxic HPUFs could be identified by the same method. The vector pCU1 was constructed by splicing *S. aureus* plasmid pCLP100 and the *E. coli* cloning vector pUC19 where protein expression is regulated under a lac promoter (Augustin et al., 1992). This shuttle plasmid pCU1 can be stably maintained in both *E. coli* and *S. aureus* and has the ability to take up to 6 kb insertion fragments. In this study, the pCU1LK contained a 45 bp linker sequence at the sites PstI to KpnI. All insertion genes ranging from 171 bp to 753 bp should allow the constructed pCU1LK derivatives to replicate freely in both hosts.

The plate-based screening assay relies on the leaky transcription of the lac promoter in a low-glucose environment without active lactose or isopropyl β-D-1-thiogalactopyranoside (IPTG) induction of gene transcription, as applied successfully in previous toxicity screenings of HPUF products (Kennell and Riezman, 1977, Mohanraj et al., 2019). Therefore, in this study, the transformants were plated on LB Amp100 plates without additional inducers or repressors. The production of potential bactericidal proteins also relied on the basal transcription under lac promoter in the pCU1LK vector. According to Kasurinen et al., basal expression was sufficient to show the toxicity of the gene products both in the plating assay and in the NGS-based assay and to provide comparable screening results from both assays (2021). Further

validation using plating-based assay on a few randomly chosen Stab21 HPUFs would provide more information regarding the induction and repression.

4.3 Conclusions, improvements, and further research

This screening study of 96 hypothetical proteins of unknown function from the Stab21 phage genome yielded fourteen potentially bacteriotoxic gene products. The screening in *E. coli* serves as a valuable starting point in pursuit of novel small antibiotic molecules that can prove useful in the treatment of current antimicrobial-resistant infections. The NGS-based method was further reinforced as a viable and useful method of high-throughput screening. The assay was adapted for the use of the pCU1LK shuttle vector, which required the addition of a dephosphorylation step after linearization of the plasmid with restriction enzymes to minimize self-ligation. Different methods of linearized vector purification were also assessed to determine the optimal conditions for the ligation reactions. It was found that both agarose gel purification and affinity-based column purification caused significant loss of DNA quality resulting in ligation mixtures with fewer correctly inserted gene fragments and transformants (data not shown). The linearized vector was therefore used in ligation reactions directly from the reaction mix after dephosphorylation.

To further validate the results from preliminary screening by NGS, expression of individual toxic candidates in a tightly regulated expression vector should be tested in both *E. coli* and *S. aureus*. An observation of reduced or completely stopped bacterial growth would confirm the toxic effect of the gene product. The screening assay in this master's thesis study provides a valuable starting point to the following steps, as it reduced the toxic candidates to be further investigated from 96 to just 14. It can be expected that several of the potentially toxic HPUFs will indeed hinder the growth of *E. coli* at least, as it has been the screening host in this study. Previous phage HPUF-screening studies have yielded between 3.1% and 13.6% of the total number of genes and between 17% and 60% of potentially toxic genes investigated with confirmed toxic activity after the initial screening assay (Mohanraj et al., 2019, Spruit et al., 2020, Kasurinen et al., 2021). If the gene products of Stab21 HPUFs follow the same patterns, 2 to 8 of the potentially toxic genes may indeed display bacteriotoxic activity in *S. aureus* in for example a growth-curve assay.

Once the toxic candidates are confirmed, their interactions with the bacterial cellular targets can be further identified and their mechanisms of toxicity can be characterised. Due to the yet to be unveiled status of these hypothetical phage proteins, it is possible to identify novel targets that are prevalent in a wide range of bacteria and new mechanisms of toxicity that have not been utilized by any existing antibiotics. Approaches to discover bacterial targets of antimicrobial molecules include genomic manipulation of the host cells and pull-down assays followed by mass spectrometry of the products (Wan et al., 2021). Structural proteomics methods such as nuclear magnetic resonance and X-ray crystallography can also be applied directly to the phage proteins of interest, to elucidate the structure and aid the interpretation of protein docking and interaction with a bacterial host (Parmar et al., 2017). To be useful therapeutically, the potential cellular targets for the elimination of pathogens should not be present in eukaryotic cells and must also be tested against these to ensure non-toxicity.

References

Augustin, J., Rosenstein, R., Wieland, B., Schneider, U., Schnell, N., Engelke, G., Entian, K. D., & Götz, F. (1992). Genetic analysis of epidermin biosynthetic genes and epidermin-negative mutants of Staphylococcus epidermidis. *European journal of biochemistry*, 204(3), 1149–1154. https://doi.org/10.1111/j.1432-1033.1992.tb16740.x

Azam, A. H., & Tanji, Y. (2019). Bacteriophage-host arm race: an update on the mechanism of phage resistance in bacteria and revenge of the phage with the perspective for phage therapy. *Applied microbiology and biotechnology*, 103(5), 2121–2131. https://doi.org/10.1007/s00253-019-09629-x

Cespedes, C., Said-Salim, B., Miller, M., Lo, S. H., Kreiswirth, B. N., Gordon, R. J., Vavagiakis, P., Klein, R. S., & Lowy, F. D. (2005). The clonality of Staphylococcus aureus nasal carriage. *The Journal of infectious diseases*, 191(3), 444–452. https://doi.org/10.1086/427240

Chaitanya K. V. (2019). Structure and Organization of Virus Genomes. *Genome and Genomics: From Archaea to Eukaryotes,* 1–30. https://doi.org/10.1007/978-981-15-0702-1_1

Coates, A. R., Halls, G., & Hu, Y. (2011). Novel classes of antibiotics or more of the same?. *British journal of pharmacology*, 163(1), 184–194. https://doi.org/10.1111/j.1476-5381.2011.01250.x

Dedrick, R. M., Guerrero-Bustamante, C. A., Garlena, R. A., Russell, D. A., Ford, K., Harris, K., Gilmour, K. C., Soothill, J., Jacobs-Sera, D., Schooley, R. T., Hatfull, G. F., & Spencer, H. (2019). Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nature medicine,* 25(5), 730–733. https://doi.org/10.1038/s41591-019-0437-z

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G., 3rd, Burland, V., Mau, B., Petrosino, J. F., Qin, X., Muzny, D. M., Ayele, M., Gibbs, R. A., Csörgo, B., Pósfai, G., Weinstock, G. M., & Blattner, F. R. (2008). The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *Journal of bacteriology*, 190(7), 2597–2606. https://doi.org/10.1128/JB.01695-07

Gordillo Altamirano, F. L., & Barr, J. J. (2019). Phage Therapy in the Postantibiotic Era. *Clinical microbiology reviews*, 32(2), e00066-18. https://doi.org/10.1128/CMR.00066-18

Hesse, S., Rajaure, M., Wall, E., Johnson, J., Bliskovsky, V., Gottesman, S., & Adhya, S. (2020). Phage Resistance in Multidrug-Resistant Klebsiella pneumoniae ST258 Evolves via Diverse Mutations That Culminate in Impaired Adsorption. *mBio*, 11(1), e02530-19. https://doi.org/10.1128/mBio.02530-19

Hover, B. M., Kim, S. H., Katz, M., Charlop-Powers, Z., Owen, J. G., Ternei, M. A., Maniko, J., Estrela, A. B., Molina, H., Park, S., Perlin, D. S., & Brady, S. F. (2018). Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. *Nature microbiology*, 3(4), 415–422. https://doi.org/10.1038/s41564-018-0110-1

Kasurinen, J., Spruit, C. M., Wicklund, A., Pajunen, M. I., & Skurnik, M. (2021). Screening of Bacteriophage Encoded Toxic Proteins with a Next Generation Sequencing-Based Assay. *Viruses,* 13(5), 750. https://doi.org/10.3390/v13050750

Kasurinen, J. (2020). Screening of bacteriophage encoded toxic proteins with an NGS based assay. Master's thesis. University of Helsinki.

Kennell, D., & Riezman, H. (1977). Transcription and translation initiation frequencies of the Escherichia coli lac operon. *Journal of molecular biology*, 114(1), 1–21. https://doi.org/10.1016/0022-2836(77)90279-0

Kim, H. Y., Banerjee, S. K., & Iyer, V. N. (1994). The incN plasmid replicon: two pathways of DNA polymerase I-independent replication. *Journal of bacteriology,* 176(24), 7735–7739. https://doi.org/10.1128/jb.176.24.7735-7739.1994

Kwon, S., Park, S., Lee, B., and Yoon, S. (2013). "In-depth analysis of interrelation between quality scores and real errors in illumina reads," *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* pp. 635-638, doi: 10.1109/EMBC.2013.6609580.

Ling, L. L., Schneider, T., Peoples, A. J., Spoering, A. L., Engels, I., Conlon, B. P., Mueller, A., Schäberle, T. F., Hughes, D. E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. A., Cohen, D. R., Felix, C. R., Fetterman, K. A., Millett, W. P., Nitti, A. G., Zullo, A. M., Chen, C., … Lewis, K. (2015). A new antibiotic kills

pathogens without detectable resistance. *Nature*, 517(7535), 455–459. https://doi.org/10.1038/nature14098

Liu, J., Dehbi, M., Moeck, G., Arhin, F., Bauda, P., Bergeron, D., Callejo, M., Ferretti, V., Ha, N., Kwan, T., McCarty, J., Srikumar, R., Williams, D., Wu, J. J., Gros, P., Pelletier, J., & DuBow, M. (2004). Antimicrobial drug discovery through bacteriophage genomics. *Nature biotechnology*, 22(2), 185–191. https://doi.org/10.1038/nbt932

Miklasińska-Majdanik M. (2021). Mechanisms of Resistance to Macrolide Antibiotics among Staphylococcus aureus. *Antibiotics* (Basel, Switzerland), 10(11), 1406. https://doi.org/10.3390/antibiotics10111406

Mohanraj, Ushanandini; Wan, Xing; Spruit, Cindy M.; Skurnik, Mikael; Pajunen, Maria I. 2019. "A Toxicity Screening Approach to Identify Bacteriophage-Encoded Anti-Microbial Proteins" *Viruses* 11, no. 11: 1057. https://doi.org/10.3390/v11111057

Monaco, M., Pimentel de Araujo, F., Cruciani, M., Coccia, E. M., & Pantosti, A. (2017). Worldwide Epidemiology and Antibiotic Resistance of Staphylococcus aureus. *Current topics in microbiology and immunology*, 409, 21–56. https://doi.org/10.1007/82_2016_3

Mulani, M. S., Kamble, E. E., Kumkar, S. N., Tawre, M. S., & Pardesi, K. R. (2019). Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. *Frontiers in microbiology*, *10*, 539. https://doi.org/10.3389/fmicb.2019.00539

Oduor, J.M.O., Kiljunen, S., Kadija, E. et al. 2019. Genomic characterization of four novel Staphylococcus myoviruses. *Arch Virol* 164, 2171–2173 https://doi.org/10.1007/s00705-019-04267-0

Oechslin F. (2018). Resistance Development to Bacteriophages Occurring during Bacteriophage Therapy. *Viruses,* 10(7), 351. https://doi.org/10.3390/v10070351

Parmar, K. M., Gaikwad, S. L., Dhakephalkar, P. K., Kothari, R., & Singh, R. P. (2017). Intriguing Interaction of Bacteriophage-Host Association: An Understanding in the Era of Omics. *Frontiers in microbiology*, 8, 559. https://doi.org/10.3389/fmicb.2017.00559

Roach, D. R., & Donovan, D. M. (2015). Antimicrobial bacteriophage-derived proteins and therapeutic applications. *Bacteriophage*, 5(3), e1062590. https://doi.org/10.1080/21597081.2015.1062590

Ruiz, N., & Silhavy, T. J. (2022). How Escherichia coli Became the Flagship Bacterium of Molecular Biology. *Journal of bacteriology*, 204(9), e0023022. https://doi.org/10.1128/jb.00230-22

Saier, M. H., Jr, & Reddy, B. L. (2015). Holins in bacteria, eukaryotes, and archaea: multifunctional xenologues with potential biotechnological and biomedical applications. *Journal of bacteriology*, 197(1), 7–17. https://doi.org/10.1128/JB.02046-14

Schoch CL, et al. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)*. baaa062. PubMed: 32761142 PMC: PMC7408187. https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2490449

Shibayama Y, Dabbs ER. Phage as a source of antibacterial genes: Multiple inhibitory products encoded by Rhodococcus phage YF1. *Bacteriophage*. 2011 Jul 1;1(4):195-197. https://doi.org/10.4161/bact.1.4.17746 PMID: 23050212; PMCID: PMC3448104.

Sharma M. (2013). Lytic bacteriophages: Potential interventions against enteric bacterial pathogens on produce. *Bacteriophage,* 3(2), e25518. https://doi.org/10.4161/bact.25518

Singh, S., Godavarthi, S., Kumar, A., & Sen, R. (2019). A mycobacteriophage genomics approach to identify novel mycobacteriophage proteins with mycobactericidal properties. *Microbiology* (Reading, England), 165(7), 722–736. https://doi.org/10.1099/mic.0.000810

Schmelcher, M., & Loessner, M. J. (2016). Bacteriophage endolysins: applications for food safety. *Current opinion in biotechnology*, 37, 76–87. https://doi.org/10.1016/j.copbio.2015.10.005

Schmelcher, M., Donovan, D. M., & Loessner, M. J. (2012). Bacteriophage endolysins as novel antimicrobials. *Future microbiology*, 7(10), 1147–1171. https://doi.org/10.2217/fmb.12.97

Spruit, C.M., Wicklund, A., Wan, X., Skurnik, M. and Pajunen, M.I., 2020. Discovery of three toxic proteins of Klebsiella phage fHe-Kpn01. *Viruses*, 12(5), p.544. https://doi.org/10.3390/v12050544

Takeuchi, F., Watanabe, S., Baba, T., Yuzawa, H., Ito, T., Morimoto, Y., Kuroda, M., Cui, L., Takahashi, M., Ankai, A., Baba, S., Fukui, S., Lee, J. C., & Hiramatsu, K. (2005). Whole-genome sequencing of staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. *Journal of bacteriology*, 187(21), 7292–7308. https://doi.org/10.1128/JB.187.21.7292-7308.2005

Uchiyama, J., Takemura-Uchiyama, I., Sakaguchi, Y., Gamoh, K., Kato, S., Daibata, M., Ujihara, T., Misawa, N., & Matsuzaki, S. (2014). Intragenus generalized transduction in Staphylococcus spp. by a novel giant phage. *The ISME journal*, 8(9), 1949–1952. https://doi.org/10.1038/ismej.2014.29

Van den Bossche, A., Ceyssens, P. J., De Smet, J., Hendrix, H., Bellon, H., Leimer, N., Wagemans, J., Delattre, A. S., Cenens, W., Aertsen, A., Landuyt, B., Minakhin, L., Severinov, K., Noben, J. P., & Lavigne, R. (2014). Systematic identification of hypothetical bacteriophage proteins targeting key protein complexes of Pseudomonas aeruginosa. *Journal of proteome research,* 13(10), 4446–4456. https://doi.org/10.1021/pr500796n

Wan, X., Hendrix, H., Skurnik, M. and Lavigne, R., 2020. Phage-based target discovery and its exploitation towards novel antibacterial molecules. *Current Opinion in Biotechnology,* 68, pp.1-7. https://doi.org/10.1016/j.copbio.2020.08.015

Watson R. (2008). Multidrug resistance responsible for half of deaths from healthcare associated infections in Europe. *BMJ (Clinical research ed.)*, 336(7656), 1266–1267. https://doi.org/10.1136/bmj.39601.623808.4E

World Health Organization. (2017, February 27). WHO publishes list of bacteria for which new antibiotics are urgently needed. *World Health Organization*. Retrieved September 9, 2022, from https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed

Appendices

Appendix 1. Stab21 HPUF-encoding genes and their primers

Restriction sites are indicated in coloured text. The NotI sites in blue, the NheI sites in gray and the KpnI sites in red.

Gene	Size (bp)	Forward Primer	Reverse Primer
g002	306	GCAGCGGCCGCatgataataatactatttacgcag	GGTGCTAGCctataagaatttctctatatgttcc
g003	297	GCAGCGGCCGCatgattgatatatacttaggag	GGTGCTAGCttaaaatatctcttctattattctt
g005	318	GCAGCGGCCGCatgatagaaattaggttagacg	GGTGCTAGCctaataatctaagtcaaaagggt
g006	291	GCAGCGGCCGCatgatagagatataccttagtg	GGTGCTAGCttacatctcctttacatactc
g008	237	GCAGCGGCCGCatggttactttaacatacactatt	GGTGGTACCctatcctacgtgccaagc
g009	345	GCAGCGGCCGCatgatagttatatatacagatgttt	GGTGCTAGCtcaatccccgcccatac
g010c	336	GCAGCGGCCGCatgaaaataaactatattccaa	GGTGCTAGCttatagaatatttataacattgtatt
g012	285	GCAGCGGCCGCatgacaaacaaaaattacttatac	GGTGCTAGCttaattcttaaccgcttctatt
g013	270	GCAGCGGCCGCatgatattagaaatagaaactaa	GGTGCTAGCttatttagtttttaattctacatta
g017	354	GCAGCGGCCGCatgaaaatgttcaaattacaaaa	GGTGCTAGCtcaatgtctgattggtct
g018	219	GCAGCGGCCGCatgaacagattagaaatagtaaa	GGTGCTAGCttatgcgtattcttcgatt
g020	171	GCAGCGGCCGCatgattacaatgacaaaaacaa	GGTGCTAGCttaaacagtttctgagttctt
g021	468	GCAGCGGCCGCatgacaaatacaatacaagcat	GGTGCTAGCctacagtgccatttttgc
g022	195	GCAGCGGCCGCatgttgaagatgaataaatac	GGTGCTAGCttacatttcttctactacataac
g024	654	GCAGCGGCCGCatgaacattaacgaatatatagg	GGTGCTAGCctacctccctaagtctttt
g026	186	GCAGCGGCCGCatgatgaacatgacaaact	GGTGCTAGCttaaaatattccattttgttttttt
g027c	234	GCAGCGGCCGCatgaaactataccaagtagaa	GGTGCTAGCttaggtattattaacaacctct
g028c	483	GCAGCGGCCGCatgaacaaagaacaagcc	GGTGCTAGCttacttattctccttgatttttttt
g029c	405	GCAGCGGCCGCatgggattagactttgaag	GGTGCTAGCttatacatttacactcatgattaa
g030c	429	GCAGCGGCCGCatggaaaattataaaaactttatt	GGTGCTAGCttatttttcctcctcttcat
g031c	246	GCAGCGGCCGCatgagatatgatattaatgaaaaat	GGTGCTAGCtcattgtgattcctcctta
g033c	429	GCAGCGGCCGCatgaatatcaaatatattgatttag	GGTGCTAGCttattcatcttcttcctcc
g034c	540	GCAGCGGCCGCatggataagataaatctcaata	GGTGCTAGCttatattaataattctttccattct
g039c	546	GCAGCGGCCGCatggaaaaaatttatatattagaag	GGTGCTAGCtcaagttaatttatcaattgaat
g040c	216	GCAGCGGCCGCatgaaaaatattattaatttttagt	GGTGCTAGCttactcccaaataccaata
g042c	735	GCAGCGGCCGCatgaacttagaaaaaagtttc	GGTGCTAGCttatctctcattatagacctc
g043c	237	GCAGCGGCCGCatggacttttaccaatttc	GGTGCTAGCttaataaccatgtttagttacc
g044c	387	GCAGCGGCCGCatgtttaaaaaagcacctc	GGTGCTAGCttactcatccttttaacgt
g045c	171	GCAGCGGCCGCatggaaaaagtaaatcatgag	GGTGCTAGCttatttagcattgtatttccatt
g046c	480	GCAGCGGCCGCatggcaaatgaaaaagaga	GGTGCTAGCtcataggtctttttctaagtca
g053c	324	GCAGCGGCCGCgtgtctaaaagaacagac	GGTGCTAGCttaaaaatacattaatttaaaaaaatc

g056c	186	GCAGCGGCCGCatggaaaaattccaagaag	GGTGCTAGCttattctatatctcctttaatttct
g061c	576	GCAGCGGCCGCatggataatttatcacattact	GGTGCTAGCctacctccttgagtaataatt
g062c	621	GCAGCGGCCGCatggtaaataaaattaacgataaa	GGTGGTACCttatccatcttgttcccc
g065c	222	GCAGCGGCCGCatgaattatttagctaaggtat	GGTGCTAGCttaattatcctcctttgaattat
g069c	189	GCAGCGGCCGCatgaaaaaaggagtatttaca	GGTGCTAGCctatcctgcatacttataatcc
g072c	225	GCAGCGGCCGCatgaataaatttaaaagatggtt	GGTGCTAGCttatttctcctctacttttaaaaa
g075c	306	GCAGCGGCCGCatggcactacttttaacat	GGTGCTAGCttacatttctccttttttctattt
g078c	183	GCAGCGGCCGCatggcatcagcaaaacaa	GGTGCTAGCttactcattaatttggtttagtttttt
g079c	216	GCAGCGGCCGCatgaaaagacaaaaaatgtttt	GGTGCTAGCttagttatcttttgttaattcttcc
g080c	207	GCAGCGGCCGCatgtcaaaacatattgaaataa	GGTGCTAGCttagaatactattttaaaagattct
g081c	330	GCAGCGGCCGCttggataaggagataaacaac	GGTGCTAGCctatgcaaatttgttaaagaca
g083c	264	GCAGCGGCCGCatgattatcgtatctttttttct	GGTGCTAGCttacttattttgtggtataatagtt
g085	276	GCAGCGGCCGCatgaaaacaaagaaagaaattaaa	GGTGGTACCtcaatccatttcacctcg
g086	408	GCAGCGGCCGCttgagtgcagaaaatattaga	GGTGCTAGCttagaatgtttctgaattttcc
g089	171	GCAGCGGCCGCgtgattttatttagcactataatc	GGTGCTAGCtcatttatttcttccttccttt
g092	339	GCAGCGGCCGCatgaatattataacgtcactat	GGTGCTAGCttatttttttatctttaaagttacttt
g093	369	GCAGCGGCCGCttgatattctctaaagataaaaaatg	GGTGCTAGCctagtcacctctactccc
g103	213	GCAGCGGCCGCatggctagaaaaaagaca	GGTGCTAGCttatatatctaatttcctacctaga
g107	456	GCAGCGGCCGCatgagtacattttggtcag	GGTGCTAGCttattgaattgtcaagtctttac
g108	192	GCAGCGGCCGCatgggtataacaatagtaaatag	GGTGCTAGCctacataaattttagtgaccaat
g109	309	GCAGCGGCCGCatgtcacaagataaattaagag	GGTGCTAGCttactttacatattcacctgtac
g131	375	GCAGCGGCCGCatgaaaaaatatagagaataccta	GGTGCTAGCttacttatccccttttcgtaa
g135	336	GCAGCGGCCGCatgtcaaataataaaaaagatatttt	GGTGCTAGCttattcttgttctccttttttcttcttc
g136	450	GCAGCGGCCGCatggaaaaaaatattagcaca	GGTGCTAGCttactgttcgtcatttttct
g141	327	GCAGCGGCCGCatggatagaaaagaagcaat	GGTGCTAGCctattcatttttttccatcttctg
g146	240	GCAGCGGCCGCgtgaatacgggagagatt	GGTGCTAGCttaaatattaactgagatactactt
g150	351	GCAGCGGCCGCatggataatttaatagataaaaaca	GGTGCTAGCttagctttcttcataaggatt
g155	258	GCAGCGGCCGCatggatattccaacaatattattt	GGTGCTAGCctactcacctactctttcat
g156	753	GCAGCGGCCGCatgggaattatagtaaactcc	GGTGCTAGCttactcataactgcttcctt
g159	309	GCAGCGGCCGCatgaagttcaatgatatttatga	GGTGCTAGCctataagaaatccttttccattttt
g166	441	GCAGCGGCCGCatgtttatttcattaaatcaagaa	GGTGCTAGCttactcaatgacaatactatcc
g169	240	GCAGCGGCCGCatggaaatggcagatttag	GGTGCTAGCctacctcctttggtctattt
g171	174	GCAGCGGCCGCatggttatacctagtattaaagc	GGTGCTAGCttactcaccatatctctcct
g172	249	GCAGCGGCCGCatggtgagtaaatttatcgg	GGTGGTACCttattcattttctttatccttaatg
g173	231	GCAGCGGCCGCatgaataaaggggaatttattat	GGTGCTAGCttagcctggttgatttact
g175	246	GCAGCGGCCGCatgataagctcatttgatagt	GGTGGTACCctatagtaaaatattgtttactgct
g176	174	GCAGCGGCCGCatggattttaatgattttataaaca	GGTGCTAGCttagtcatttctttttctcctt
g177	294	GCAGCGGCCGCatgactaaagaaacaaatgtac	GGTGCTAGCttaaaatgcttcatctgtcaa
g179	366	GCAGCGGCCGCatggatatactaattattcattataa	GGTGCTAGCttataacattaagtcttcatttaat
g181	276	GCAGCGGCCGCatgcctatggacttattaac	GGTGCTAGCttaagaaaatgaaagaagatttatt

g182	312	GCAGCGGCCGCatgattaatatgagtaaagaaac	GGTGCTAGCCtataattgtaacttatgatagttaa
g183	348	GCAGCGGCCGCatgagagaagagttaaaacc	GGTGCTAGCttatttttcctccttttgtaac
g185	177	GCAGCGGCCGCatgaatgagtggtatgct	GGTGCTAGCttatctctccttatcaaattctt
g187	291	GCAGCGGCCGCatgaagcagagagattttg	GGTGCTAGCttaaatatctaatttctcataacaat
g188	285	GCAGCGGCCGCatgaacaaagcagtagaa	GGTGCTAGCctactttataaaacctttaagttc
g190	303	GCAGCGGCCGCatgaatggtattattgtattttac	GGTGCTAGCttattgactcatctcctctaa
g191	402	GCAGCGGCCGCatggtaattgcgtttttttat	GGTGCTAGCtcactccttattaagttcaatt
g192	234	GCAGCGGCCGCatggaatttatagataaaaataatgt	GGTGCTAGCtcatagtatgtcctccttttt
g194	318	GCAGCGGCCGCgtggagaaattcaaaggt	GGTGCTAGCttatttccctccttcaatct
g195	228	GCAGCGGCCGCatggaatatttatttttatttatagg	GGTGCTAGCttaaaagaataaaatcttaatttctt
g196	177	GCAGCGGCCGCatgaaacattttattttaattttagg	GGTGCTAGCttaatttctactaaacatacttcc
g198	315	GCAGCGGCCGCatgaaagtagaatcaatagca	GGTGCTAGCttatttttcctccttaaaatatctt
g199	678	GCAGCGGCCGCgtgtctaataaaactattacaaa	GGTGCTAGCttaatttttaatgatacctactaat
g201	222	GCAGCGGCCGCatgaattatgaagaggtact	GGTGCTAGCttaaaataaaatagctcctgc
g202	198	GCAGCGGCCGCatgaattatagagattttattacaga	GGTGCTAGCttataacccctccgttgt
g204	306	GCAGCGGCCGCatgtatcctgaaatagatgt	GGTGCTAGCtcattttgttgatagctcc
g206	393	GCAGCGGCCGCatggtaaaattagataaatacttaa	GGTGCTAGCttagtattctccttctgttatt
g208	243	GCAGCGGCCGCatgatttataaaatatcaaaacataa	GGTGCTAGCctatggctgtaaccattc
g209	390	GCAGCGGCCGCatgattatagataaattaaatggag	GGTGCTAGCctatttctctccttttaattcttt
g210	195	GCAGCGGCCGCatgagtaatagttgggaaaaa	GGTGCTAGCttatttatctgctacatactcat
g211	294	GCAGCGGCCGCatgatgaatggaaaacaaat	GGTGCTAGCttacatacctttcacatagtc
g212	309	GCAGCGGCCGCatgaaaaaactattaatattatttac	GGTGCTAGCttaatctcctttatatattaattcat
g213	237	GCAGCGGCCGCatgtatatattagaaagaacaattag	GGTGCTAGCtcataagtcattctcccac
g215	192	GCAGCGGCCGCatgataaatatagaacatgattatac	GGTGCTAGCttaccatcgttcaatagatac
g216	351	GCAGCGGCCGCatgaatgctaggaaagca	GGTGCTAGCttaccaactaatgtatataataggt

Appendix 2. Ligation-joint sequences read coverages

Gene	Joint sequence	Ligations reads	Plasmid reads
g002	ATGCCTGCAG**GCGGCCGC**ATGATAATAA	0	0
	TTATTATCAT**GCGGCCGC**CTGCAGGCAT	0	23
	GATCCTCTAGA**GCTAGC**CTATAAGAATT	0	0
	AATTCTTATAG**GCTAGC**TCTAGAGGATC	0	0
g003	ATGCCTGCAG**GCGGCCGC**ATGATTGATA	1730	18305
	TATCAATCAT**GCGGCCGC**CTGCAGGCAT	16	14114
	GATCCTCTAGA**GCTAGC**TTAAAATATCT	2206	16031
	AGATATTTTAA**GCTAGC**TCTAGAGGATC	118	16608
g005	ATGCCTGCAG**GCGGCCGC**ATGATAGAAA	2011	13464
	TTTCTATCAT**GCGGCCGC**CTGCAGGCAT	24	10057
	GATCCTCTAGA**GCTAGC**CTAATAATCTA	2170	12502
	TAGATTATTAG**GCTAGC**TCTAGAGGATC	88	11375
g006	ATGCCTGCAG**GCGGCCGC**ATGATAGAGA	1443	14081
	TCTCTATCAT**GCGGCCGC**CTGCAGGCAT	9	11108
	GATCCTCTAGA**GCTAGC**TTACATCTCCT	1969	14576
	AGGAGATGTAA**GCTAGC**TCTAGAGGATC	153	13754
	ATGCCTGCAG**GCGGCCGC**ATGGTTACTT	342	1058

g008	AAGTAACCAT**GCGGCCGC**CTGCAGGCAT	380	948
	AATTCGAGCTC**GGTACC**CTATCCTACGT	2805	1014
	ACGTAGGATAG**GGTACC**GAGCTCGAATT	1619	1205
g009	ATGCCTGCAG**GCGGCCGC**ATGATAGTTA	1984	19842
	TAACTATCAT**GCGGCCGC**CTGCAGGCAT	12	15330
	GATCCTCTAGA**GCTAGC**TCAATCCCCGC	2452	20046
	GCGGGGATTGA**GCTAGC**TCTAGAGGATC	117	16929
g010c	ATGCCTGCAG**GCGGCCGC**ATGAAAATAA	1380	17371
	TTATTTTCAT**GCGGCCGC**CTGCAGGCAT	5	13694
	GATCCTCTAGA**GCTAGC**TTATAGAATAT	1427	15422
	ATATTCTATAA**GCTAGC**TCTAGAGGATC	52	15546
g012	ATGCCTGCAG**GCGGCCGC**ATGACAAACA	1994	18602
	TGTTTGTCAT**GCGGCCGC**CTGCAGGCAT	11	10041
	GATCCTCTAGA**GCTAGC**TTAATTCTTAA	2100	15189
	TTAAGAATTAA**GCTAGC**TCTAGAGGATC	125	16569
g013	ATGCCTGCAG**GCGGCCGC**ATGATATTAG	893	14591
	CTAATATCAT**GCGGCCGC**CTGCAGGCAT	8	11736
	GATCCTCTAGA**GCTAGC**TTATTTAGTTT	1043	12546
	AAACTAAATAA**GCTAGC**TCTAGAGGATC	91	12334
g017	ATGCCTGCAG**GCGGCCGC**ATGAAAATGT	1640	11934
	ACATTTTCAT**GCGGCCGC**CTGCAGGCAT	11	8705
	GATCCTCTAGA**GCTAGC**TCAATGTCTGA	1797	11823
	TCAGACATTGA**GCTAGC**TCTAGAGGATC	58	10723
g018	ATGCCTGCAG**GCGGCCGC**ATGAACAGAT	0	0
	ATCTGTTCAT**GCGGCCGC**CTGCAGGCAT	0	0
	GATCCTCTAGA**GCTAGC**TTATGCGTATT	0	0
	AATACGCATAA**GCTAGC**TCTAGAGGATC	0	0
g020	ATGCCTGCAG**GCGGCCGC**ATGATTACAA	183	13515
	TTGTAATCAT**GCGGCCGC**CTGCAGGCAT	15	11481
	GATCCTCTAGA**GCTAGC**TTAAACAGTTT	307	12171
	AAACTGTTTAA**GCTAGC**TCTAGAGGATC	48	14301
g021	ATGCCTGCAG**GCGGCCGC**ATGACAAATA	1264	7618
	TATTTGTCAT**GCGGCCGC**CTGCAGGCAT	7	4485
	GATCCTCTAGA**GCTAGC**CTACAGTGCCA	895	6800
	TGGCACTGTAG**GCTAGC**TCTAGAGGATC	143	5766
g022	ATGCCTGCAG**GCGGCCGC**ATGTTGAAGA	170	13327
	TCTTCAACAT**GCGGCCGC**CTGCAGGCAT	9	12253
	GATCCTCTAGA**GCTAGC**TTACATTTCTT	473	13093
	AAGAAATGTAA**GCTAGC**TCTAGAGGATC	120	14251
g024	ATGCCTGCAG**GCGGCCGC**ATGAACATTA	1510	5473
	TAATGTTCAT**GCGGCCGC**CTGCAGGCAT	9	4309
	GATCCTCTAGA**GCTAGC**CTACCTCCCTA	1491	5106
	TAGGGAGGTAG**GCTAGC**TCTAGAGGATC	63	5354
g026	ATGCCTGCAG**GCGGCCGC**ATGATGAACA	144	9578
	TGTTCATCAT**GCGGCCGC**CTGCAGGCAT	2	8994
	GATCCTCTAGA**GCTAGC**TTAAAATATTC	313	8224
	GAATATTTTAA**GCTAGC**TCTAGAGGATC	97	11175
g027c	ATGCCTGCAG**GCGGCCGC**ATGAAACTAT	299	5930
	ATAGTTTCATGCGGCCGCCTGCAGGCAT	2	6819
	GATCCTCTAGA**GCTAGC**TTAGGTATTAT	845	5288
	ATAATACCTAAGCTAGCTCTAGAGGATC	357	7588
g028c	ATGCCTGCAG**GCGGCCGC**ATGAACAAAG	921	8265
	CTTTGTTCATGCGGCCGCCTGCAGGCAT	2	9143
	GATCCTCTAGA**GCTAGC**TTACTTATTCT	1032	8548
	AGAATAAGTAAGCTAGCTCTAGAGGATC	122	10759
g029c	ATGCCTGCAG**GCGGCCGC**ATGGGATTAG	747	4673
	CTAATCCCATGCGGCCGCCTGCAGGCAT	3	5114
	GATCCTCTAGA**GCTAGC**TTATACATTTA	981	4046
	TAAATGTATAAGCTAGCTCTAGAGGATC	89	5773

g030c	ATGCCTGCAG**GCGGCCGC**ATGGAAAATT	867	4584
	AATTTTCCATGCGGCCGCCTGCAGGCAT	7	5282
	GATCCTCTAGA**GCTAGC**TTATTTTTCCT	975	4678
	AGGAAAAATAAGCTAGCTCTAGAGGATC	85	6289
g031c	ATGCCTGCAG**GCGGCCGC**ATGAGATATG	568	3391
	CATATCTCATGCGGCCGCCTGCAGGCAT	3	3936
	GATCCTCTAGA**GCTAGC**TCATTGTGATT	1170	3363
	AATCACAATGAGCTAGCTCTAGAGGATC	103	4924
g033c	ATGCCTGCAG**GCGGCCGC**ATGAATATCA	878	11930
	TGATATTCATGCGGCCGCCTGCAGGCAT	3	11898
	GATCCTCTAGA**GCTAGC**TTATTCATCTT	794	9566
	AAGATGAATAAGCTAGCTCTAGAGGATC	76	16237
g034c	ATGCCTGCAG**GCGGCCGC**ATGGATAAGA	703	6452
	TCTTATCCATGCGGCCGCCTGCAGGCAT	3	6920
	GATCCTCTAGA**GCTAGC**TTATATTAATA	891	5370
	TATTAATATAAGCTAGCTCTAGAGGATC	441	8702
g039c	ATGCCTGCAG**GCGGCCGC**ATGGAAAAAA	1086	9984
	TTTTTTCCATGCGGCCGCCTGCAGGCAT	5	11647
	GATCCTCTAGA**GCTAGC**TCAAGTTAATT	1104	8761
	AATTAACTTGAGCTAGCTCTAGAGGATC	55	15136
g040c	ATGCCTGCAG**GCGGCCGC**ATGAAAAATA	296	7810
	TATTTTTCATGCGGCCGCCTGCAGGCAT	1	10260
	GATCCTCTAGA**GCTAGC**TTACTCCCAAA	653	7974
	TTTGGGAGTAAGCTAGCTCTAGAGGATC	99	12182
g042c	ATGCCTGCAG**GCGGCCGC**ATGAACTTAG	1160	7959
	CTAAGTTCATGCGGCCGCCTGCAGGCAT	3	9039
	GATCCTCTAGA**GCTAGC**TTATCTCTCAT	1901	8065
	ATGAGAGATAAGCTAGCTCTAGAGGATC	43	11325
g043c	ATGCCTGCAG**GCGGCCGC**ATGGACTTTT	353	5639
	AAAAGTCCATGCGGCCGCCTGCAGGCAT	3	6674
	GATCCTCTAGA**GCTAGC**TTAATAACCAT	926	4945
	ATGGTTATTAAGCTAGCTCTAGAGGATC	117	8861
g044c	ATGCCTGCAG**GCGGCCGC**ATGTTTAAAA	761	4666
	TTTTAAACATGCGGCCGCCTGCAGGCAT	4	5404
	GATCCTCTAGA**GCTAGC**TTACTCATCCT	1044	5080
	AGGATGAGTAAGCTAGCTCTAGAGGATC	100	6231
g045c	ATGCCTGCAG**GCGGCCGC**ATGGAAAAAG	79	7532
	CTTTTTCCATGCGGCCGCCTGCAGGCAT	3	8772
	GATCCTCTAGA**GCTAGC**TTATTTAGCAT	323	7217
	ATGCTAAATAAGCTAGCTCTAGAGGATC	71	10313
g046c	ATGCCTGCAG**GCGGCCGC**ATGGCAAATG	619	4333
	CATTTGCCATGCGGCCGCCTGCAGGCAT	3	5196
	GATCCTCTAGA**GCTAGC**TCATAGGTCTT	728	4359
	AAGACCTATGAGCTAGCTCTAGAGGATC	716	5626
g053c	ATGCCTGCAG**GCGGCCGC**GTGTCTAAAA	864	5144
	TTTTAGACACGCGGCCGCCTGCAGGCAT	5	6568
	GATCCTCTAGA**GCTAGC**TTAAAAATACA	1399	4628
	TGTATTTTTAAGCTAGCTCTAGAGGATC	84	6527
g056c	ATGCCTGCAG**GCGGCCGC**ATGGAAAAAT	150	7190
	ATTTTTCCATGCGGCCGCCTGCAGGCAT	2	9591
	GATCCTCTAGA**GCTAGC**TTATTCTATAT	400	6410
	ATATAGAATAAGCTAGCTCTAGAGGATC	174	10619
g061c	ATGCCTGCAG**GCGGCCGC**ATGGATAATT	1559	14144
	AATTATCCAT**GCGGCCGC**CTGCAGGCAT	16	10584
	GATCCTCTAGA**GCTAGC**CTACCTCCTTG	1338	12183
	CAAGGAGGTAG**GCTAGC**TCTAGAGGATC	208	12048
g062c	ATGCCTGCAG**GCGGCCGC**ATGGTAAATA	779	5
	TATTTACCAT**GCGGCCGC**CTGCAGGCAT	1000	0
	AATTCGAGCTC**GGTACC**TTATCCATCTT	0	0

	AAGATGGATAA**GGTACC**GAGCTCGAATT	0	2
g065c	ATGCCTGCAG**GCGGCCGC**ATGAATTATT	412	14899
	AATAATTCAT**GCGGCCGC**CTGCAGGCAT	4	14530
	GATCCTCTAGA**GCTAGC**TTAATTATCCT	654	13964
	AGGATAATTAA**GCTAGC**TCTAGAGGATC	81	15130
g069c	ATGCCTGCAG**GCGGCCGC**ATGAAAAAAG	126	10685
	CTTTTTTCAT**GCGGCCGC**CTGCAGGCAT	12	11462
	GATCCTCTAGA**GCTAGC**CTATCCTGCAT	334	10922
	ATGCAGGATAG**GCTAGC**TCTAGAGGATC	60	12853
g072c	ATGCCTGCAG**GCGGCCGC**ATGAATAAAT	400	9677
	ATTTATTCAT**GCGGCCGC**CTGCAGGCAT	15	9153
	GATCCTCTAGA**GCTAGC**TTATTTCTCCT	440	9529
	AGGAGAAATAA**GCTAGC**TCTAGAGGATC	611	10772
g075c	ATGCCTGCAG**GCGGCCGC**ATGGCACTAC	494	7644
	GTAGTGCCAT**GCGGCCGC**CTGCAGGCAT	5	6296
	GATCCTCTAGA**GCTAGC**TTACATTTCTC	1984	6379
	GAGAAATGTAA**GCTAGC**TCTAGAGGATC	226	7114
g078c	ATGCCTGCAG**GCGGCCGC**ATGGCATCAG	45	2562
	CTGATGCCAT**GCGGCCGC**CTGCAGGCAT	0	2299
	GATCCTCTAGA**GCTAGC**TTACTCATTAA	276	2319
	TTAATGAGTAA**GCTAGC**TCTAGAGGATC	160	2929
g079c	ATGCCTGCAG**GCGGCCGC**ATGAAAAGAC	128	3063
	GTCTTTTCAT**GCGGCCGC**CTGCAGGCAT	2	2689
	GATCCTCTAGA**GCTAGC**TTAGTTATCTT	572	3156
	AAGATAACTAA**GCTAGC**TCTAGAGGATC	60	3836
g080c	ATGCCTGCAG**GCGGCCGC**ATGTCAAAAC	64	4665
	GTTTTGACAT**GCGGCCGC**CTGCAGGCAT	2	4205
	GATCCTCTAGA**GCTAGC**TTAGAATACTA	358	4167
	TAGTATTCTAA**GCTAGC**TCTAGAGGATC	48	4954
g081c	ATGCCTGCAG**GCGGCCGC**TTGGATAAGG	349	42
	CCTTATCCAA**GCGGCCGC**CTGCAGGCAT	2	35
	GATCCTCTAGA**GCTAGC**CTATGCAAATT	928	35
	AATTTGCATAG**GCTAGC**TCTAGAGGATC	45	40
g083c	ATGCCTGCAG**GCGGCCGC**ATGATTATCG	196	5441
	CGATAATCAT**GCGGCCGC**CTGCAGGCAT	2	4797
	GATCCTCTAGA**GCTAGC**TTACTTATTTT	1088	4839
	AAAATAAGTAA**GCTAGC**TCTAGAGGATC	48	4993
g085	ATGCCTGCAG**GCGGCCGC**ATGAAAACAA	320	1
	TTGTTTTCAT**GCGGCCGC**CTGCAGGCAT	39	1
	AATTCGAGCTC**GGTACC**TCAATCCATTT	0	0
	AAATGGATTGA**GGTACC**GAGCTCGAATT	0	0
g086	ATGCCTGCAG**GCGGCCGC**TTGAGTGCAG	787	15794
	CTGCACTCAA**GCGGCCGC**CTGCAGGCAT	20	11996
	GATCCTCTAGA**GCTAGC**TTAGAATGTTT	1068	16931
	AAACATTCTAA**GCTAGC**TCTAGAGGATC	107	16324
g089	ATGCCTGCAG**GCGGCCGC**GTGATTTTAT	91	16926
	ATAAAATCAC**GCGGCCGC**CTGCAGGCAT	2	16141
	GATCCTCTAGA**GCTAGC**TCATTTATTTC	232	15550
	GAAATAAATGA**GCTAGC**TCTAGAGGATC	118	19076
g092	ATGCCTGCAG**GCGGCCGC**ATGAATATTA	1042	15222
	TAATATTCAT**GCGGCCGC**CTGCAGGCAT	8	11823
	GATCCTCTAGA**GCTAGC**TTATTTTTTAT	1648	11679
	ATAAAAAATAA**GCTAGC**TCTAGAGGATC	120	12184
g093	ATGCCTGCAG**GCGGCCGC**TTGATATTCT	575	13067
	AGAATATCAA**GCGGCCGC**CTGCAGGCAT	0	11160
	GATCCTCTAGA**GCTAGC**CTAGTCACCTC	876	14390
	GAGGTGACTAG**GCTAGC**TCTAGAGGATC	50	14443
g103	ATGCCTGCAG**GCGGCCGC**ATGGCTAGAA	295	14649
	TTCTAGCCAT**GCGGCCGC**CTGCAGGCAT	3	14742

	GATCCTCTAGA**GCTAGC**TTATATATCTA	578	13232
	TAGATATATAA**GCTAGC**TCTAGAGGATC	136	14589
g107	ATGCCTGCAG**GCGGCCGC**ATGAGTACAT	423	5214
	ATGTACTCAT**GCGGCCGC**CTGCAGGCAT	6	4472
	GATCCTCTAGA**GCTAGC**TTATTGAATTG	731	4762
	CAATTCAATAA**GCTAGC**TCTAGAGGATC	25	4733
g108	ATGCCTGCAG**GCGGCCGC**ATGGGTATAA	196	10591
	TTATACCCAT**GCGGCCGC**CTGCAGGCAT	1	11567
	GATCCTCTAGA**GCTAGC**CTACATAAATT	282	10029
	AATTTATGTAG**GCTAGC**TCTAGAGGATC	37	12880
g109	ATGCCTGCAG**GCGGCCGC**ATGTCACAAG	1462	11628
	CTTGTGACAT**GCGGCCGC**CTGCAGGCAT	8	9564
	GATCCTCTAGA**GCTAGC**TTACTTTACAT	1883	10600
	ATGTAAAGTAA**GCTAGC**TCTAGAGGATC	70	12109
g131	ATGCCTGCAG**GCGGCCGC**ATGAAAAAAT	2010	19918
	ATTTTTTCAT**GCGGCCGC**CTGCAGGCAT	19	17931
	GATCCTCTAGA**GCTAGC**TTACTTATCCC	1687	19345
	GGGATAAGTAA**GCTAGC**TCTAGAGGATC	88	20080
g135	ATGCCTGCAG**GCGGCCGC**ATGTCAAATA	2214	21754
	TATTTGACAT**GCGGCCGC**CTGCAGGCAT	9	17892
	GATCCTCTAGA**GCTAGC**TTATTCTTGTT	2045	17129
	AACAAGAATAA**GCTAGC**TCTAGAGGATC	211	23892
g136	ATGCCTGCAG**GCGGCCGC**ATGGAAAAAA	1044	10515
	TTTTTTCCAT**GCGGCCGC**CTGCAGGCAT	2	8105
	GATCCTCTAGA**GCTAGC**TTACTGTTCGT	1294	9623
	ACGAACAGTAA**GCTAGC**TCTAGAGGATC	114	11642
g141	ATGCCTGCAG**GCGGCCGC**ATGGATAGAA	1840	17450
	TTCTATCCAT**GCGGCCGC**CTGCAGGCAT	9	16618
	GATCCTCTAGA**GCTAGC**CTATTCATTTT	1621	16339
	AAAATGAATAG**GCTAGC**TCTAGAGGATC	81	18377
g146	ATGCCTGCAG**GCGGCCGC**GTGAATACGG	764	22415
	CCGTATTCAC**GCGGCCGC**CTGCAGGCAT	7	21605
	GATCCTCTAGA**GCTAGC**TTAAATATTAA	877	19599
	TTAATATTTAA**GCTAGC**TCTAGAGGATC	75	22639
g150	ATGCCTGCAG**GCGGCCGC**ATGGATAATT	2145	16698
	AATTATCCAT**GCGGCCGC**CTGCAGGCAT	4	14616
	GATCCTCTAGA**GCTAGC**TTAGCTTTCTT	2049	15909
	AAGAAAGCTAA**GCTAGC**TCTAGAGGATC	56	18003
g155	ATGCCTGCAG**GCGGCCGC**ATGGATATTC	806	12255
	GAATATCCAT**GCGGCCGC**CTGCAGGCAT	8	11693
	GATCCTCTAGA**GCTAGC**CTACTCACCTA	953	12237
	TAGGTGAGTAG**GCTAGC**TCTAGAGGATC	122	12197
g156	ATGCCTGCAG**GCGGCCGC**ATGGGAATTA	1958	5361
	TAATTCCCAT**GCGGCCGC**CTGCAGGCAT	6	4467
	GATCCTCTAGA**GCTAGC**TTACTCATAAC	2134	4467
	GTTATGAGTAA**GCTAGC**TCTAGAGGATC	78	5644
g159	ATGCCTGCAG**GCGGCCGC**ATGAAGTTCA	1799	5820
	TGAACTTCAT**GCGGCCGC**CTGCAGGCAT	10	5161
	GATCCTCTAGA**GCTAGC**CTATAAGAAAT	1609	5529
	ATTTCTTATAG**GCTAGC**TCTAGAGGATC	49	5817
g166	ATGCCTGCAG**GCGGCCGC**ATGTTTATTT	1205	14716
	AAATAAACAT**GCGGCCGC**CTGCAGGCAT	3	13389
	GATCCTCTAGA**GCTAGC**TTACTCAATGA	1201	13822
	TCATTGAGTAA**GCTAGC**TCTAGAGGATC	68	16807
g169	ATGCCTGCAG**GCGGCCGC**ATGGAAATGG	522	13832
	CCATTTCCAT**GCGGCCGC**CTGCAGGCAT	8	12031
	GATCCTCTAGA**GCTAGC**CTACCTCCTTT	545	12422
	AAAGGAGGTAG**GCTAGC**TCTAGAGGATC	85	15266
	ATGCCTGCAG**GCGGCCGC**ATGGTTATAC	141	14296

g171	GTATAACCAT**GCGGCCGC**CTGCAGGCAT	7	15531
	GATCCTCTAGA**GCTAGC**TTACTCACCAT	288	13772
	ATGGTGAGTAA**GCTAGC**TCTAGAGGATC	127	18288
g172	ATGCCTGCAG**GCGGCCGC**ATGGTGAGTA	2563	3951
	TACTCACCAT**GCGGCCGC**CTGCAGGCAT	1953	3861
	AATTCGAGCTC**GGTACC**TTATTCATTTT	1876	3654
	AAAATGAATAA**GGTACC**GAGCTCGAATT	1872	4162
g173	ATGCCTGCAG**GCGGCCGC**ATGAATAAAG	430	9036
	CTTTATTCAT**GCGGCCGC**CTGCAGGCAT	3	9107
	GATCCTCTAGA**GCTAGC**TTAGCCTGGTT	791	9070
	AACCAGGCTAA**GCTAGC**TCTAGAGGATC	108	10610
g175	ATGCCTGCAG**GCGGCCGC**ATGATAAGCT	2655	1577
	AGCTTATCAT**GCGGCCGC**CTGCAGGCAT	1501	1506
	AATTCGAGCTC**GGTACC**CTATAGTAAAA	1505	1287
	TTTTACTATAG**GGTACC**GAGCTCGAATT	1504	1581
g176	ATGCCTGCAG**GCGGCCGC**ATGGATTTTA	203	13572
	TAAAATCCAT**GCGGCCGC**CTGCAGGCAT	13	13653
	GATCCTCTAGA**GCTAGC**TTAGTCATTTC	310	11411
	GAAATGACTAA**GCTAGC**TCTAGAGGATC	142	19780
g177	ATGCCTGCAG**GCGGCCGC**ATGACTAAAG	939	5024
	CTTTAGTCAT**GCGGCCGC**CTGCAGGCAT	3	4536
	GATCCTCTAGA**GCTAGC**TTAAAATGCTT	1486	4815
	AAGCATTTTAA**GCTAGC**TCTAGAGGATC	48	4907
g179	ATGCCTGCAG**GCGGCCGC**ATGGATATAC	2048	16889
	GTATATCCAT**GCGGCCGC**CTGCAGGCAT	7	13147
	GATCCTCTAGA**GCTAGC**TTATAACATTA	1832	14793
	TAATGTTATAA**GCTAGC**TCTAGAGGATC	75	15998
g181	ATGCCTGCAG**GCGGCCGC**ATGCCTATGG	972	8760
	CCATAGGCAT**GCGGCCGC**CTGCAGGCAT	7	7904
	GATCCTCTAGA**GCTAGC**TTAAGAAAATG	1386	7246
	CATTTTCTTAA**GCTAGC**TCTAGAGGATC	52	8649
g182	ATGCCTGCAG**GCGGCCGC**ATGATTAATA	1575	11392
	TATTAATCAT**GCGGCCGC**CTGCAGGCAT	10	8969
	GATCCTCTAGA**GCTAGC**CTATAATTGTA	1691	9725
	TACAATTATAG**GCTAGC**TCTAGAGGATC	54	12104
g183	ATGCCTGCAG**GCGGCCGC**ATGAGAGAAG	2133	16463
	CTTCTCTCAT**GCGGCCGC**CTGCAGGCAT	11	14642
	GATCCTCTAGA**GCTAGC**TTATTTTTCCT	2114	15280
	AGGAAAAATAA**GCTAGC**TCTAGAGGATC	187	16520
g185	ATGCCTGCAG**GCGGCCGC**ATGAATGAGT	161	12612
	ACTCATTCAT**GCGGCCGC**CTGCAGGCAT	7	13467
	GATCCTCTAGA**GCTAGC**TTATCTCTCCT	344	11823
	AGGAGAGATAA**GCTAGC**TCTAGAGGATC	205	18012
g187	ATGCCTGCAG**GCGGCCGC**ATGAAGCAGA	1979	2574
	TCTGCTTCAT**GCGGCCGC**CTGCAGGCAT	27	1955
	GATCCTCTAGA**GCTAGC**TTAAATATCTA	2000	1748
	TAGATATTTAA**GCTAGC**TCTAGAGGATC	130	2518
g188	ATGCCTGCAG**GCGGCCGC**ATGAACAAAG	1783	14223
	CTTTGTTCAT**GCGGCCGC**CTGCAGGCAT	11	11812
	GATCCTCTAGA**GCTAGC**CTACTTTATAA	1562	11609
	TTATAAAGTAG**GCTAGC**TCTAGAGGATC	77	12860
g190	ATGCCTGCAG**GCGGCCGC**ATGAATGGTA	1893	8627
	TACCATTCAT**GCGGCCGC**CTGCAGGCAT	14	7671
	GATCCTCTAGA**GCTAGC**TTATTGACTCA	1927	8067
	TGAGTCAATAA**GCTAGC**TCTAGAGGATC	144	8981
g191	ATGCCTGCAG**GCGGCCGC**ATGGTAATTG	1389	15268
	CAATTACCAT**GCGGCCGC**CTGCAGGCAT	11	13432
	GATCCTCTAGA**GCTAGC**TCACTCCTTAT	1271	14498
	ATAAGGAGTGA**GCTAGC**TCTAGAGGATC	64	15746

g192	ATGCCTGCAG**GCGGCCGC**ATGGAATTTA	649	10453
	TAAATTCCAT**GCGGCCGC**CTGCAGGCAT	11	10588
	GATCCTCTAGA**GCTAGC**TCATAGTATGT	756	10773
	ACATACTATGA**GCTAGC**TCTAGAGGATC	66	12407
g194	ATGCCTGCAG**GCGGCCGC**GTGGAGAAAT	1498	14979
	ATTTCTCCAC**GCGGCCGC**CTGCAGGCAT	4	12749
	GATCCTCTAGA**GCTAGC**TTATTTCCCTC	1296	10810
	GAGGGAAATAA**GCTAGC**TCTAGAGGATC	112	16970
g195	ATGCCTGCAG**GCGGCCGC**ATGGAATATT	576	21518
	AATATTCCAT**GCGGCCGC**CTGCAGGCAT	3	21476
	GATCCTCTAGA**GCTAGC**TTAAAAGAATA	634	20210
	TATTCTTTTAA**GCTAGC**TCTAGAGGATC	55	24508
g196	ATGCCTGCAG**GCGGCCGC**ATGAAACATT	113	22625
	AATGTTTCAT**GCGGCCGC**CTGCAGGCAT	2	26451
	GATCCTCTAGA**GCTAGC**TTAATTTCTAC	297	22396
	GTAGAAATTAA**GCTAGC**TCTAGAGGATC	84	28687
g198	ATGCCTGCAG**GCGGCCGC**ATGAAAGTAG	1272	16580
	CTACTTTCAT**GCGGCCGC**CTGCAGGCAT	9	14317
	GATCCTCTAGA**GCTAGC**TTATTTTTCCT	918	14888
	AGGAAAAATAA**GCTAGC**TCTAGAGGATC	34	15988
g199	ATGCCTGCAG**GCGGCCGC**GTGTCTAATA	1585	31107
	TATTAGACAC**GCGGCCGC**CTGCAGGCAT	3	27280
	GATCCTCTAGA**GCTAGC**TTAATTTTTAA	2520	30016
	TTAAAAATTAA**GCTAGC**TCTAGAGGATC	51	28325
g201	ATGCCTGCAG**GCGGCCGC**ATGAATTATG	129	5882
	CATAATTCAT**GCGGCCGC**CTGCAGGCAT	1	5945
	GATCCTCTAGA**GCTAGC**TTAAAATAAAA	400	5747
	TTTTATTTTAA**GCTAGC**TCTAGAGGATC	40	7118
g202	ATGCCTGCAG**GCGGCCGC**ATGAATTATA	98	2961
	TATAATTCAT**GCGGCCGC**CTGCAGGCAT	1	3102
	GATCCTCTAGA**GCTAGC**TTATAACCCCT	574	2962
	AGGGGTTATAA**GCTAGC**TCTAGAGGATC	119	3790
g204	ATGCCTGCAG**GCGGCCGC**ATGTATCCTG	608	9444
	CAGGATACAT**GCGGCCGC**CTGCAGGCAT	1	7892
	GATCCTCTAGA**GCTAGC**TCATTTTGTTG	1376	8468
	CAACAAAATGA**GCTAGC**TCTAGAGGATC	78	9925
g206	ATGCCTGCAG**GCGGCCGC**ATGGTAAAAT	797	13275
	ATTTTACCAT**GCGGCCGC**CTGCAGGCAT	5	11383
	GATCCTCTAGA**GCTAGC**TTAGTATTCTC	1403	12359
	GAGAATACTAA**GCTAGC**TCTAGAGGATC	137	13683
g208	ATGCCTGCAG**GCGGCCGC**ATGATTTATA	157	4288
	TATAAATCAT**GCGGCCGC**CTGCAGGCAT	1	4228
	GATCCTCTAGA**GCTAGC**CTATGGCTGTA	523	4290
	TACAGCCATAG**GCTAGC**TCTAGAGGATC	48	5572
g209	ATGCCTGCAG**GCGGCCGC**ATGATTATAG	269	3734
	CTATAATCAT**GCGGCCGC**CTGCAGGCAT	4	2978
	GATCCTCTAGA**GCTAGC**CTATTTCTCTC	620	3646
	GAGAGAAATAG**GCTAGC**TCTAGAGGATC	652	4115
g210	ATGCCTGCAG**GCGGCCGC**ATGAGTAATA	56	5881
	TATTACTCAT**GCGGCCGC**CTGCAGGCAT	0	5983
	GATCCTCTAGA**GCTAGC**TTATTTATCTG	388	5843
	CAGATAAATAA**GCTAGC**TCTAGAGGATC	183	6909
g211	ATGCCTGCAG**GCGGCCGC**ATGATGAATG	493	7520
	CATTCATCAT**GCGGCCGC**CTGCAGGCAT	0	6642
	GATCCTCTAGA**GCTAGC**TTACATACCTT	1219	7327
	AAGGTATGTAA**GCTAGC**TCTAGAGGATC	87	8579
g212	ATGCCTGCAG**GCGGCCGC**ATGAAAAAAC	509	2261
	GTTTTTTCAT**GCGGCCGC**CTGCAGGCAT	62	1753
	GATCCTCTAGA**GCTAGC**TTAATCTCCTT	996	1831

	AAGGAGATTAA**GCTAGC**TCTAGAGGATC	221	2100
g213	ATGCCTGCAG**GCGGCCGC**ATGTATATAT	974	6394
	ATATATACAT**GCGGCCGC**CTGCAGGCAT	179	6264
	GATCCTCTAGA**GCTAGC**TCATAAGTCAT	1562	6269
	ATGACTTATGA**GCTAGC**TCTAGAGGATC	254	8408
g215	ATGCCTGCAG**GCGGCCGC**ATGATAAATA	216	2012
	TATTTATCAT**GCGGCCGC**CTGCAGGCAT	68	1869
	GATCCTCTAGA**GCTAGC**TTACCATCGTT	755	1857
	AACGATGGTAA**GCTAGC**TCTAGAGGATC	125	2756
g216	ATGCCTGCAG**GCGGCCGC**ATGAATGCTA	589	2226
	TAGCATTCAT**GCGGCCGC**CTGCAGGCAT	64	1729
	GATCCTCTAGA**GCTAGC**TTACCAACTAA	1079	1937
	TTAGTTGGTAA**GCTAGC**TCTAGAGGATC	265	2040

Appendix 3. Total and relative ligation-joint reads

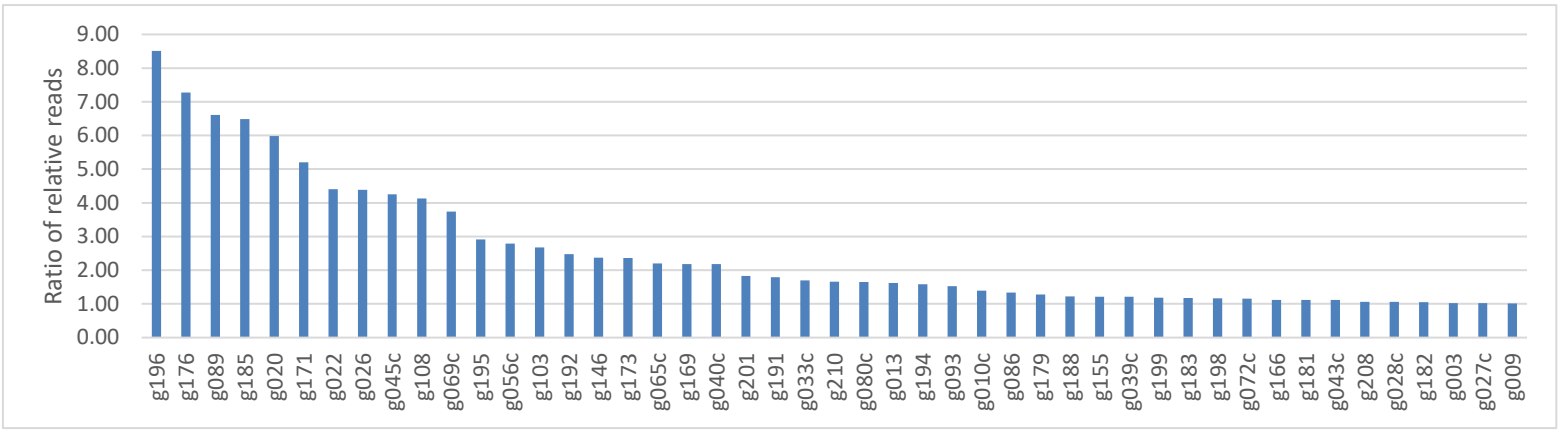
Genes are grouped by gene pool. Ratios are calculated as relative plasmid reads divided by relative ligation reads.

Gene	Total ligation reads	Relative ligation reads (%)	Total plasmid reads	Relative plasmid reads (%)	Ratio
g002	0	0.0	23	0.0	-
g003	4070	9.8	65058	10.1	1.0
g005	4293	10.3	47398	7.3	0.7
g006	3574	8.6	53519	8.3	1.0
g008	5146	12.4	4225	0.7	0.1
g009	4565	11.0	72147	11.2	1.0
g010c	2864	6.9	62033	9.6	1.4
g012	4230	10.2	60401	9.3	0.9
g013	2035	4.9	51207	7.9	1.6
g017	3506	8.4	43185	6.7	0.8
g018	0	0.0	0	0.0	-
g020	553	1.3	51468	8.0	6.0
g021	2309	5.6	24669	3.8	0.7
g022	772	1.9	52924	8.2	4.4
g024	3073	7.4	20242	3.1	0.4
g026	556	1.3	37971	5.9	4.4
g027c	1503	5.3	25625	5.4	1.0
g028c	2077	7.3	36715	7.8	1.1
g029c	1820	6.4	19606	4.1	0.6
g030c	1934	6.8	20833	4.4	0.6
g031c	1844	6.5	15614	3.3	0.5
g033c	1751	6.2	49631	10.5	1.7

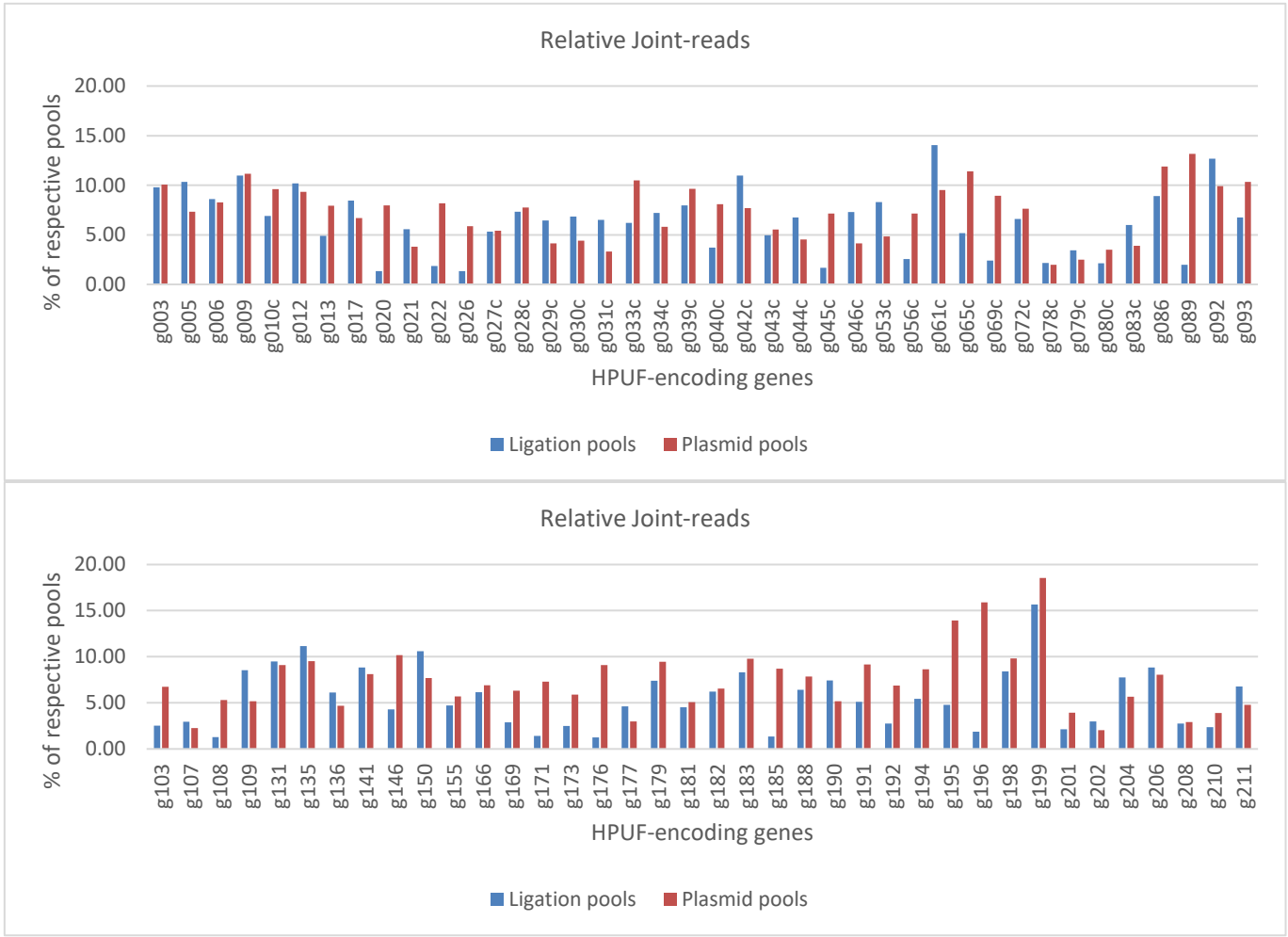
g034c	2038	7.2	27444	5.8	0.8
g039c	2250	8.0	45528	9.6	1.2
g040c	1049	3.7	38226	8.1	2.2
g042c	3107	11.0	36388	7.7	0.7
g043c	1399	4.9	26119	5.5	1.1
g044c	1909	6.7	21381	4.5	0.7
g045c	476	1.7	33834	7.2	4.3
g046c	2066	7.3	19514	4.1	0.6
g053c	2352	8.3	22867	4.8	0.6
g056c	726	2.6	33810	7.1	2.8
g061c	3121	14.0	48959	9.5	0.7
g062c	1779	8.0	7	0.0	0.0
g065c	1151	5.2	58523	11.4	2.2
g069c	532	2.4	45922	8.9	3.7
g072c	1466	6.6	39131	7.6	1.2
g075c	2709	12.2	27433	5.3	0.4
g078c	481	2.2	10109	2.0	0.9
g079c	762	3.4	12744	2.5	0.7
g080c	472	2.1	17991	3.5	1.7
g081c	1324	6.0	152	0.0	0.0
g083c	1334	6.0	20070	3.9	0.7
g085	359	1.6	2	0.0	0.0
g086	1982	8.9	61045	11.9	1.3
g089	443	2.0	67693	13.2	6.6
g092	2818	12.7	50908	9.9	0.8
g093	1501	6.8	53060	10.3	1.5
g103	1012	2.5	57212	6.7	2.7
g107	1185	3.0	19181	2.3	0.8
g108	516	1.3	45067	5.3	4.1
g109	3423	8.5	43901	5.2	0.6
g131	3804	9.5	77274	9.1	1.0
g135	4479	11.2	80667	9.5	0.9
g136	2454	6.1	39885	4.7	0.8
g141	3551	8.8	68784	8.1	0.9
g146	1723	4.3	86258	10.2	2.4
g150	4254	10.6	65226	7.7	0.7
g155	1889	4.7	48382	5.7	1.2

g156	4176	10.4	19939	2.4	0.2
g159	3467	8.6	22327	2.6	0.3
g166	2477	6.2	58734	6.9	1.1
g169	1160	2.9	53551	6.3	2.2
g171	563	1.4	61887	7.3	5.2
g172	8264	15.5	15628	2.4	0.2
g173	1332	2.5	37823	5.9	2.4
g175	7165	13.4	5951	0.9	0.1
g176	668	1.2	58416	9.1	7.3
g177	2476	4.6	19282	3.0	0.6
g179	3962	7.4	60827	9.5	1.3
g181	2417	4.5	32559	5.1	1.1
g182	3330	6.2	42190	6.6	1.1
g183	4445	8.3	62905	9.8	1.2
g185	717	1.3	55914	8.7	6.5
g187	4136	7.7	8795	1.4	0.2
g188	3433	6.4	50504	7.9	1.2
g190	3978	7.4	33346	5.2	0.7
g191	2735	5.1	58944	9.2	1.8
g192	1482	2.8	44221	6.9	2.5
g194	2910	5.4	55508	8.6	1.6
g195	1268	4.8	87712	13.9	2.9
g196	496	1.9	100159	15.9	8.5
g198	2233	8.4	61773	9.8	1.2
g199	4159	15.7	116728	18.5	1.2
g201	570	2.1	24692	3.9	1.8
g202	792	3.0	12815	2.0	0.7
g204	2063	7.8	35729	5.7	0.7
g206	2342	8.8	50700	8.1	0.9
g208	729	2.7	18378	2.9	1.1
g209	1545	5.8	14473	2.3	0.4
g210	627	2.4	24616	3.9	1.7
g211	1799	6.8	30068	4.8	0.7
g212	1788	6.7	7945	1.3	0.2
g213	2969	11.2	27335	4.3	0.4
g215	1164	4.4	8494	1.3	0.3
g216	1997	7.5	7932	1.3	0.2

Appendix 4. Relative joint-sequence reads ratios for non-toxic genes

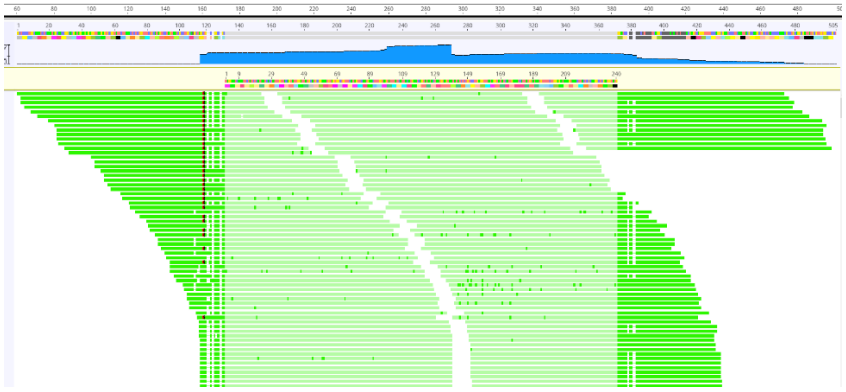


Appendix 5. Relative joint-sequence reads of non-toxic genes

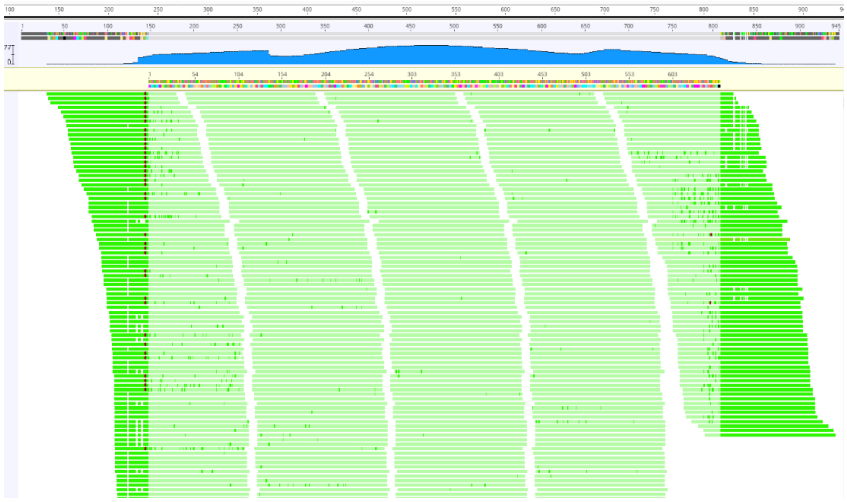


Appendix 6. Mapping of NGS reads to Stab21 hypothetical genes

A: g008

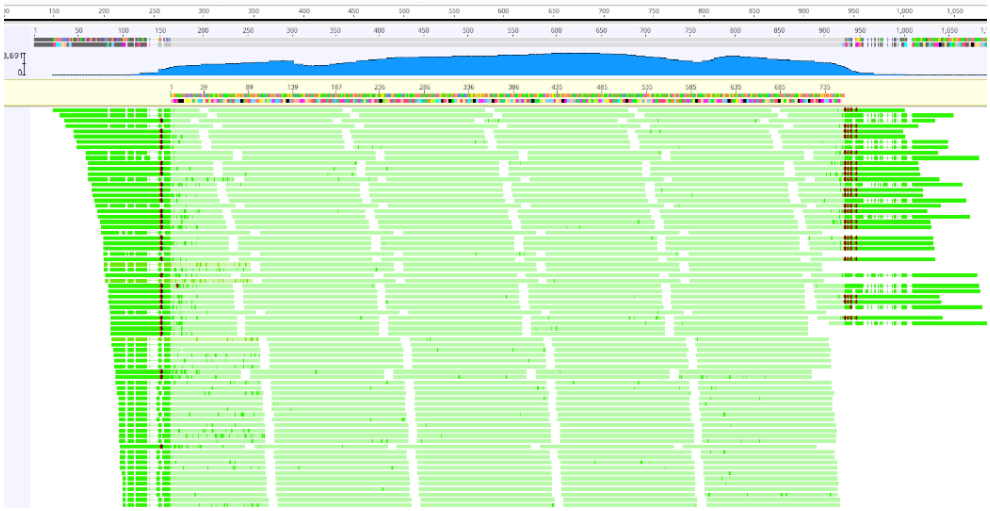


B: g024



C: g085

D: g156



E: g159

F: g175