Transcriptional analysis of *Bdellovibrio bacteriovorus* responses to different prey and environments

A thesis submitted by

Carey Lambert

As part of the requirement for the degree of:

MRes: Bioinformatics Scientist

This research was carried out under the supervision of Dr Adam Blanchard and Professor Liz Sockett

Abstract

Bdellovibrio bacteriovorus is a small, highly motile Gram-negative bacterium that preys upon other Gram-negative bacteria. It does this by burrowing through the outer layers of the prey cell and establishing itself in the periplasm, before consuming the prey contents and using these for growth as a filament. When the prey contents are depleted, this filament septates to produce new *B. bacteriovorus* progeny which burst out of the prey shell and go on to attack further prey. Its prey includes pathogens of plants, animals and humans, including multidrug-resistant pathogens which are emerging as a major health threat. Thus it has potential as a novel therapeutic to overcome the lack of new antibiotics. This prospect is particularly attractive as genetic resistance to *B. bacteriovorus* predation has not been demonstrated, rather a plastic resistance to a sub-population is seen, which when recovered remains as susceptible to predation as the parent population.

There are many stages to this complex predatory lifecycle: swimming or gliding to search out potential prey, attachment to and detecting suitable prey, formation of, and entry into, a pore in the outer layers which is then re-sealed. Then killing and rounding of the prey is followed by staged degradation of the prey contents, then growth and division of the predator ultimately leading to new prey emerging. Further, Host Independent (HI) mutants are capable of growth in rich nutrients in the lab, with divergent morphologies observed. Some transcriptional studies have proved invaluable as a tool for studying some of these processes, but a complete lifecycle study is lacking. Here, we present a high-resolution transcriptional profile throughout the predation cycle giving insights to all of these various stages.

For *B. bacteriovorus* to fulfil its promise as a novel antimicrobial agent, more needs to be understood about predation outwith the paradigm laboratory conditions. In order to address this, predation carried out on a multidrug-resistant clinical isolate of *Serratia marcescens* was subjected to transcriptional analysis, including predation by a *B. bacteriovorus* mutated by deletion of the global regulator DgcC, a strain incapable of HI growth.

Cluster analysis provides an unbiased means of ordering data by calculating distances between datapoints in *n*-dimensions, allowing grouping of gene expression from different experiments. Here, I develop a pipeline for the analysis of *B. bacteriovorus* RNA-Seq data with cluster analyses to bring further insights to both unpublished work from our laboratory and by re-analysing datasets of published work by other groups.

These analyses discover that groups of genes are tightly sequentially regulated throughout the predatory cycle, giving insight to their functions. They show that predation upon *Serratia* is significantly different from that of predation on *E. coli,* with different transcriptional profiles throughout the predation cycles. The response of *B. bacteriovorus* to exposure to pooled human serum seems to be one of protection from the antimicrobial elements in serum rather than metabolism of potential nutrients in the medium. The response to nutrient broth and non-prey Gram-positive *Staphylococcus aureus* surprisingly includes genes which are specific to Gram-negative prey modification, suggesting that the obligate predator *B. bacteriovorus* co-regulates predation and nutrient utilisation pathways. Many groups of genes are identified by analyses of mutant transcription as important in various regulatory pathways.

Along with insights into the predation process and condition responses, the gene clusters generated in this project by novel re-analyses of data identify many targets for future projects to better understand the predation process and how *B. bacteriovorus* reacts in

more clinically relevant conditions; a prerequisite for the fulfilment of its promise as a potential novel antimicrobial therapy.

Acknowledgements

Thanks to Simona Huwiler and David Negus for carrying out Experiments 1-4, for providing fastq files for this analysis and for initial analysis of differentially regulated genes for these experiments. Thanks to Adam Blanchard for supervising the project and providing invaluable support throughout the apprenticeship. Thanks to Liz Sockett for supervision and supporting the apprenticeship, providing the time to complete it and feedback on the write up. Thanks to lab members of C15 for valuable discussions and support throughout the apprenticeship, especially Emma Banks, Callum Clark and Jess Tyson.

Contents

Introduction

B. bacteriovorus is a Gram-negative predatory bacterium of the class Oligoflexia, (formerly classified as a Deltaproteobacterium, grouped with other predatory bacteria), which preys upon other Gram-negative bacteria (Stolp and Petzold, 1962, Stolp and Starr, 1963). It belongs to a diverse group termed BALOs for *B. bacteriovorus* and like organisms, which include the alphaproteobacterium *Micavibrio* along with the sea dwelling *Halobacteriovorax* and other bdellovibrios characterised by their ability to prey upon other Gram-negative organisms. *B. bacteriovorus* has a biphasic predatory lifecycle with a small (1 x 0.45 μm), highly motile attack phase which swims until it encounters a suitable prey, in which it then establishes itself and enters a growth phase (Figure 1). It grows at the expense of the prey before replicating and dividing, breaking out of the prey to repeat the predatory lifecycle again.

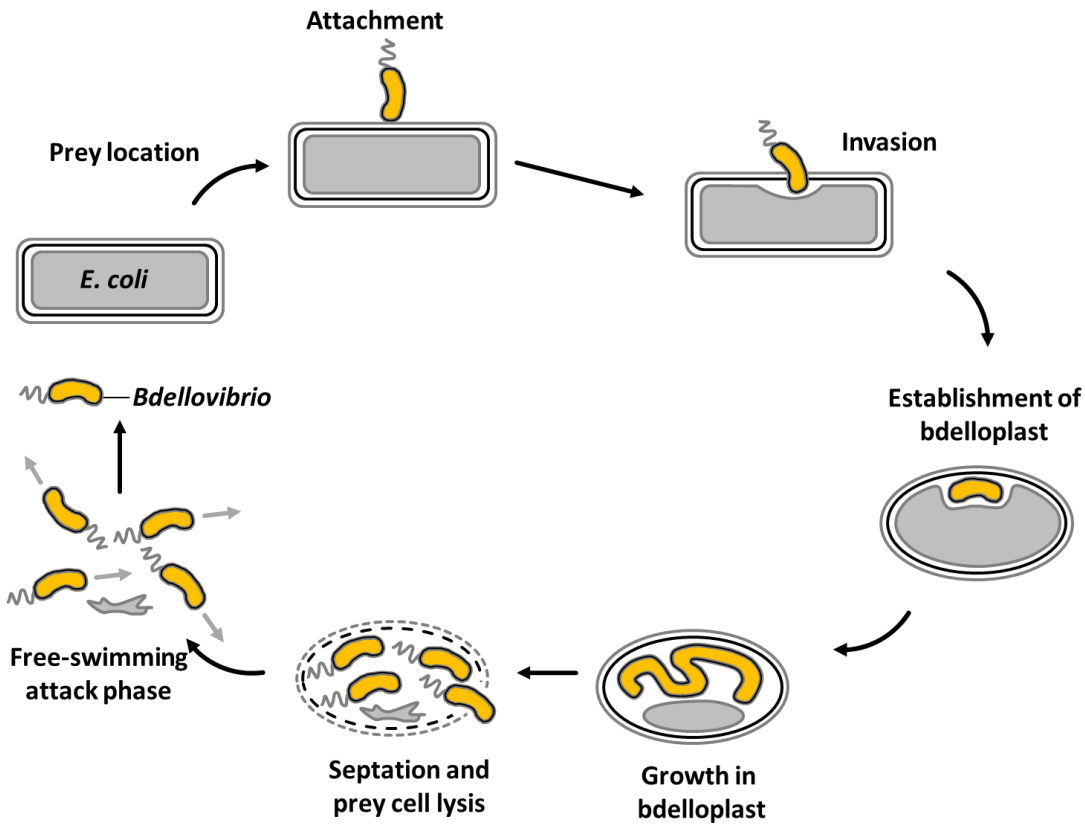

Figure 1- Predatory lifecycle of *B. bacteriovorus.*

B. bacteriovorus first attaches to a suitable prey cell, penetrates the outer layers by forming a pore of slightly smaller diameter than itself, through which it squeezes into the prey periplasm. It then establishes itself in the periplasm by re-sealing the pore by an unknown mechanism, and rounding up the prey by the action of carboxypeptidases (Lerner et al., 2012) to form a bdelloplast. It rapidly kills the prey and begins to digest the prey content and uses this to grow as a filament within the periplasm. Prey content then appears to be degraded at apparently different times, with biochemical studies suggesting that RNA and DNA are degraded early after prey killing (Hespell et al., 1975, Matin and Rittenberg, 1972). It would appear, therefore, that successive rounds of digestive enzymes must be secreted into the prey to digest RNA, DNA, protein, carbohydrates and lipids. There are a large number of predicted secreted proteases in the *B. bacteriovorus* genome (Rendulic et al., 2004) many of which must be non-specific and thus logically these have to be secreted after other degradative enzymes in order to not digest these enzymes before they have completed their activity. It was seen that at 30 minutes post-mixing of *B. bacteriovorus* and prey *E. coli*, that only a subset of genes were expressed at this timepoint (Lambert et al., 2010a). The products of these digestions are taken up by the *B. bacteriovorus* and used as building blocks for its own cell contents.

B. bacteriovorus HD100 (the Type strain and first genome sequenced (Rendulic et al., 2004)) has a surprisingly large (3,782,950 bp) genome for such a small, obligate intracellular bacterium. The same is true for the closely related strains Tiberius (3,988,594 bp (Hobley et al., 2012b)) and 109J (3,830,430 (Wurtzel et al., 2010)), suggesting that *B. bacteriovorus* does not have a reduced genome common to most intracellular bacteria. *B. bacteriovorus* HD100, the focus of this study, has a single circular genome with 50% G+C content and with 3,584 predicted open reading frames encoding a near full complement of housekeeping genes common to free-living bacteria,

but lacking some required for amino acid metabolism. The genome is also predicted to encode an uncommonly large number of degradative enzymes and transmembrane transport systems. The large genome presents *B. bacteriovorus* with a challenge as it needs to construct (usually) 4-5 such genomes in one replication cycle, but receives material from the prey genome equivalent to just one of these, so must use other prey material to build the others.

Genome replication takes place concurrently with growth throughout the filament until the prey contents are exhausted, whereupon the filament undergoes synchronous division to yield either odd or even progeny numbers of attack phase cells. The number of progeny is dependent upon the size of prey, with up to 100 recorded for large filamentous prey cells (Thomashow and Rittenberg, 1979). The newly divided progeny then grow flagella, enzymatically digest the remaining bdelloplast cell wall with a modified lysozyme (Harding et al., 2020) to burst out of the cell. If in liquid, they then swim at great speed (up to 100 cell lengths per second (Lambert et al., 2006)), if on a surface they use gliding motility (Lambert et al., 2011) to encounter further prey cells and repeat the predation cycle.

B. bacteriovorus are obligate predators, unable to grow on standard laboratory media. However, rare ($\sim 1 \times 10^{-7}$) mutants arise which are capable of growing Host Independently (HI) in the absence of prey (Reiner and Shilo, 1969). These arise mostly, but not exclusively, from mutations in the *hit* locus *bd0108,* which is associated with pili regulation (Capeness et al., 2013, Barel and Jurkevitch, 2001). HI cells seem to mimic growth within the periplasm in that they adopt a large variety of morphologies, with some attack phase cells, some elongated filaments and some rounded cells resembling bdelloplasts. HI populations retain their ability to prey, but generally with lower efficiency compared to attack phase cells (likely reflecting the low proportion of attack phase cells in the population).

B. bacteriovorus preys upon a wide range of Gram-negative organisms with different prey ranges for different strains (Dashiff et al., 2011, Dashiff and Kadouri, 2011, Jurkevitch et al., 2000). Little is known of the details of this specificity as the prey receptor has not been identified, with both prey OMPs and lipids thought to be involved in recognition (Varon and Shilo, 1969). Similarly, the mode of *B. bacteriovorus* attachment is unclear, with first a non-specific attachment which can also occur on inanimate surfaces and unsuitable prey (e.g. Gram-positive bacteria), followed by irreversible attachment and entry to suitable prey. This step is known to involve pili (Evans et al., 2007), but other factors are also involved. Included in its wide range of prey are pathogens of plants, animals and humans (Fratamico and Whiting, 1995, Negus et al., 2017, Dashiff et al., 2011) and thus *B. bacteriovorus* has great potential for use as a novel antimicrobial therapy. This is particularly timely as widespread resistance to antibiotics in Gram-negative pathogens is emerging as an ever-growing problem (Livermore, 2009). Of particular advantage is the fact that genetic resistance by prey to *B. bacteriovorus* predation has never been observed, rather a plastic resistance by a subpopulation is observed and when recovered, these are as susceptible to predation as before (Kadouri et al., 2013, Shemesh and Jurkevitch, 2004). These favourable traits have elicited a spate of promising feasibility studies for use of *B. bacteriovorus* as treatment (Atterbury et al., 2011, Shatzkes et al., 2015, Shatzkes et al., 2016, Willis et al., 2016). However, for *B. bacteriovorus* to be used as a therapeutic, more research is needed into its responses to different prey and environments, particularly in clinically relevant conditions.

Transcriptional analyses of *B. bacteriovorus* predation has been a very productive avenue of research. In the earliest genome-wide study, gene expression was compared

between attack phase cells, cells 30 minutes into predation (in bdelloplasts) and host-independent expression by microarray (Lambert et al., 2010a). By comparing the difference between attack phase and 30 minutes post-prey interaction, and also comparing attack phase to HI growth, then subtracting the former from the latter, 240 genes specific to predation were identified (the "predatosome"). This provided an invaluable source for identifying genes involved in predatory processes, including exported nucleases (Lambert and Sockett, 2013), peptidoglycan hydrolases (Lerner et al., 2012) and L,D-transpeptidases involved in sculpting the bdelloplast wall (Kuru et al., 2017) amongst many others still the subject of current studies. This pioneering study established transcriptional profiling as an important tool in understanding the *B. bacteriovorus* predation process.

The second transcriptional study of *B. bacteriovorus* prey interaction was an attempt to profile the prey response to attack by *B. bacteriovorus* using macroarrays (Lambert et al., 2010b). This showed that a variety of shock responses were induced by the attacked prey at 15 minutes post-mixing, but that these were general in nature and not a specific attempt to resist predation, further supporting the idea that prey are not capable of resisting *B. bacteriovorus* predation, and thus supporting their potential use as a novel therapy. Only a sub-population of prey evade killing, but this resistance is plastic, with the resulting rescued population as susceptible to predation after growth as the parent population (Shemesh and Jurkevitch, 2004). This lack of genetic resistance to *B. bacteriovorus* predation gives its potential use as a therapeutic an advantage over phage therapy, where resistance is readily selected for.

These initial transcriptional studies were very informative, but also limited to a couple of conditions and so the aims of this study is to expand upon these, using data gathered by colleagues (with advice from myself). These studies expanded the conditions tested to cover the whole of the predation lifecycle and different prey and conditions, to include predation in more clinically relevant conditions with pathogens in pooled human serum. My aim was to expand upon their initial analyses to use cluster analysis to pinpoint specific groups of genes involved in different conditions at specific timepoints. Hierarchical clustering computes the closest neighbours of data and clusters these iteratively to construct a dendrogram which informs overall patterns of relationships between groups of data (or expression in the case of RNA-Seq). This gives a good oversight into how many logical groups the data may be arranged. *K*-means clustering calculates the distance of data in *n* dimensions and clusters iteratively in an unbiased way. By combining these techniques and expected outcomes (by, for example, testing different combinations and checking if the resulting groups of genes seem realistic) the aim of this project is to identify transcriptional patterns in ways not previously applied to *B. bacteriovorus* transcriptional analyses. In addition to re-analysing the data in this way, the aim was to compare across experiments to find groups of genes in common or different in different conditions and timepoints. To this end, I also gathered data from published RNA-Seq experiments to re-analyse by cluster analysis (which had not been carried out in the published work) and to further compare these datasets with our own.

Experiments used in this study

Several transcriptional studies have been carried out by us and other research groups and analysis of these is the subject of this work.

Experiment 1- Predation by Wild- Type *B. bacteriovorus HD100* on *E. coli* K12 MG1655 in buffer throughout the predation cycle.

The hypothesis for Experiment 1 was that analysing expression across the whole predatory lifecycle could identify groups of genes important at each stage and give insights into processes at each timepoint throughout. By using the paradigm system of *B. bacteriovorus* HD100 Type strain preying upon *E. coli* K12 MG1655 Type strain in optimised laboratory conditions in buffer the aim was to set a baseline of transcription which can be compared against other datasets and conditions. By expanding upon initial analyses and using cluster analyses, the aim was to identify groups of genes temporally and possibly functionally related.

Experiment 1 was conducted in our laboratory by Dr Simona Huwiler with help from myself. It consists of predation by *B. bacteriovorus* HD100 (the Type strain) preying upon *E. coli* K12 MG1655 (also Type strain) throughout the predation cycle at high temporal resolution, with samples at every 15 minutes for the first hour, then at each hour up to and including 5 hours. The predator was in excess (~2-4: 1 predator to prey ratio) in order to achieve a semi-synchronous infection with all prey nearly simultaneously invaded. Microscopic observations confirmed that at 30 minutes >95% of the prey were rounded up with a *B. bacteriovorus* predator within, confirming that the invasion was a synchronous as possible to achieve. Due to the stochastic nature of attachment and invasion, there is inevitably some variation in what stage of predation each interaction is, and the excess of predators means that there are always a pool of attack phase *B. bacteriovorus* in the background. The experiment was carried out in Ca/HEPES buffer which provides no nutrients, such that all of the nutrients available to the *B. bacteriovorus* were from the prey (which had also been resuspended in Ca/HEPES buffer after overnight growth to stationary phase in YT medium).

The main aim of Experiment 1 was to establish a full transcriptional profile throughout the whole predation cycle, but also to examine transcription towards the end of the cycle, which had previously not been investigated, hence the timepoints at 4 and 5 hours. The whole predation cycle typically lasts 3 hours for smaller prey and 4 hours for average to large sized *E. coli*, so the 5 hour timepoint mostly includes the post-exit metabolism of the newly formed attack phase cells. It is historically more common to use larger but morphologically more size-variable *E. coli* S17-1 as prey in laboratory experiments, but for this experiment, the strain K12 MG1655 was chosen as it has a more uniform size and thus the later stages of predation are likely to continue more uniformly in closer synchrony to determine genes associated with the final stages of predation, including prey bursting and exit.

Experiment 2- Predation by Wild-Type *B. bacteriovorus* HD100 on *Serratia marcescens* in buffer throughout the predation cycle.

The aim of Experiment 2 was to determine any differences between the lab paradigm of predation on *E. coli* prey and predation upon a clinical isolate of the pathogen *Serratia marcescens.*

Experiment 2 was also conducted in our laboratory by David Negus, Simona Huwiler and myself. It consists of predation by *B. bacteriovorus* HD100 on a multi-drug resistant clinical isolate of *Serratia marcescens* in buffer. *Serratia* is a Gram negative Gammaproteobacterium like *E. coli*, but is an opportunistic pathogen of the family Yersinaceae, and the isolate used in this experiment is a recent clinical isolate resistant to a wide spectrum of antibiotics. *Serratia* causes hospital-acquired infections including catheter-associated bacteremia, urinary tract infections, and wound infections and thus

may be a potential target for *B. bacteriovorus* therapy. A secondary aim was to determine if the antibiotic resistance of this organism had any effect on predation by *B. bacteriovorus*. The conditions of this experiment were similar to those of Experiment 1 in order to compare predatory growth inside, and initial interaction with, a lab strain of *E. coli* with that on a more clinically relevant Gammaproteobacterium. As *B. bacteriovorus* has the potential to be used as a novel therapeutic, details of predation on such strains is important. Timepoints of 0, 2, 4, 6 and 24 hours were taken in order to get snapshots of predation throughout the predation cycle and also include the 24 hour timepoint to determine if the prey could survive predation and/or recover.

Experiment 3- Predation by Wild-Type *B. bacteriovorus* HD100 on *Serratia marcescens* in pooled human serum samples throughout the predation cycle.

The aim of Experiment 3 was to determine if there is a difference between predation in the lab paradigm conditions of buffer and more clinically relevant media such as pooled human serum.

Experiment 3 was also conducted in our laboratory by David Negus, Simona Huwiler and myself. This was identical to Experiment 2, with predation on the same clinical isolate of *Serratia marcescens,* with timepoints at 0, 2, 4, 6 and 12 hours, only instead of buffer, the medium was pooled samples of human serum. Serum is both a very challenging medium for bacteria, with high levels of antimicrobial agents (such as complement), but is also a very nutrient rich medium, so the response of both predator (which cannot metabolise media for growth in attack phase but can be stressed or killed by antimicrobials) and prey (which can metabolise or be stressed and killed by serum antimicrobials) in conditions which are more realistic to clinical settings is important to observe. The predation inside the *Serratia* prey may also be affected by it being stressed by the external conditions.

Experiment 4- Predation by *B. bacteriovorus* HD100 mutant *ΔdgcC* on *Serratia marcescens* in buffer and pooled human serum samples throughout the predation cycle.

The mutant strain *B. bacteriovorus* HD100 *ΔdgcC* has a deleted global regulator gene *dgcC*, the product of which is a GGDEF cyclic-di-GMP producing enzyme which controls the ability to switch to Host-Independent (HI) growth (Hobley et al., 2012a). The hypothesis is that the strain's inability to switch on HI growth may result in a dedicated predator, which would prey effectively and not be able to persist without prey; a desirable outcome for use as a clinical therapy. Alternatively, the lack of an important global regulator may hinder normal functional regulation and therefore predation. This could manifest itself as a change in global expression of growth or surface attachment or monitoring genes which usually permit the HI growth pattern which is a surface associated and cell-cell communicating lifestyle. Experiment 4 was also conducted in our laboratory by David Negus, Simona Huwiler and myself. Experiment 4 is essentially identical to Experiments 2 and 3 except predation throughout the timepoints is with the mutant *ΔdgcC* strain in both buffer and serum, except due to costs, the 4 hour timepoint was left out and samples were at 0, 2, 6 and 12 hours.

Experiment 5- Interaction of *B. bacteriovorus* HD100 with the Gram-positive *Staphylococcus aureus*.

The aim of Experiment 5 was to analyse the interaction between *B. bacteriovorus* and *Staphylococcus aureus* biofilms in more detail.

B. bacteriovorus is an intracellular predator of other Gram-negative bacteria, growing within the prey periplasm, and cannot prey upon Gram-positive bacteria which lack periplasms. However Im and co-workers noted that Gram-positive *Staphylococcus aureus* biofilms were dispersed to some degree by the presence of *B. bacteriovorus* with concomitant reduction in *Staphylococcus* numbers and increased viability of *B. bacteriovorus* (Im et al., 2018). To study this interaction, they isolated RNA from biofilms of *S. aureus* after incubation with *B. bacteriovorus*, and *B. bacteriovorus* controls in buffer alone. Their conclusions were that *B. bacteriovorus* produce specific proteases to break down the biofilm matrix and disperse the biofilm and that *B. bacteriovorus* express a range of metabolic genes analogous to those produced in response to nutrients, in order that they can use the nutrients released by the degrading biofilm. Here, by analysing *B. bacteriovorus* gene transcription in their data in comparison to Experiment 1, the intention is to further compare the link between genes expressed in predation and those expressed in response to *S. aureus* biofilms.

Experiment 6- *B. bacteriovorus* response to nutrients

The aim of Experiment 6 was to determine the transcriptional response of *B. bacteriovorus* to nutrients in the absence of prey.

B. bacteriovorus attack phase cells are incapable of DNA replication and cell division without acquiring a mutation (usually, but not exclusively, at the *bd0108* locus (Capeness et al., 2013)) to turn them in to host independent (HI) mutants. As such, they are incapable of growth in normal lab media, but several reports suggest that they benefit from external nutrients without growing. Experiment 6 set out to address this by comparing transcription from total RNA from attack phase cells in nutrient-free Ca/HEPES buffer with those suspended in nutrient rich broth (1 x Nutrient Broth; NB). Again, further understanding of this response can be achieved by comparing to expression during predation in Experiment 1.

Experiment 7- *B. bacteriovorus* response to Diffusible Signal Factor

Diffusible Signal Factor (DSF) is a quorum sensing molecule originally discovered being produced by the plant pathogen *Xanthamonas campestris,* and later found to be produced by other Gram-negative bacteria, that is toxic to *B. bacteriovorus*. To determine the mechanism of this toxicity, global transcription was compared between attack phase cells and intraperiplasmic *B. bacteriovorus* with and without DSF in buffer (at levels comparable to those produced by bacteria). This experiment was carried out on the strain *B. bacteriovorus* 109J, but this could be mapped onto strain HD100 in order to analyse any overlap between response to DSF and predation.

Experiment 8- Prey interactions of different predation-deficient host independent *B. bacteriovorus* mutants

The aims of Experiment 8 was to compare transcription patterns of HI mutants stalled at different regulatory points in the early part of predation.

Our laboratory has researched several potential genes involved in early predation signalling as the predator recognises, binds to, and starts to enter prey cells. Firstly,

Milner and co-workers (Milner et al., 2014) discovered that MglA, a dynamic pilus pole-defining protein in *Myxococcus* (Leonardy et al., 2010), had been evolutionarily repurposed by *B. bacteriovorus* to define the single, prey-interacting pole at which pili form. The mutant generated in this research, Δ*mglA*, formed fewer pili compared to control strains, was incapable of predation, and could only be rescued as a host independent (HI) strain. The strain was incapable of irreversible attachment, instead only transiently attaching to prey in the same way as attachment to inert surfaces or unsuitable prey. This suggests that the strain is defective in actual attachment to prey rather than specifically being involved in predatory signalling.

We identified the major pilin which is the main structural unit of pili and is essential for predation, as being encoded by *bd1290* (Evans et al., 2007). Downstream and apparently in an operon with this is the gene *bd1291*, annotated as *pilG* and containing 2 TPR domains, which are often involved in protein-protein interactions and are domains found in another protein interacting with MglA; Bd2492 (Milner et al., 2014). Deletion of *bd1291* was also only possible by rescuing HI strains and resulted in a mutant strain which formed an irreversible attachment to prey, but was unable to enter the prey. The attachment seemed to be strong, with the outer membranes of prey and predators seemingly fused as determined by electron microscopy, and prolonged incubation increased the numbers of predators attached to prey, indicating few predators dropped off when attached. However, no further predatory activity at all was shown in these strains, with the prey cell not rounding up and the predator not entering the prey in any cases. This phenotype suggests that correct signalling has occurred up to the point of prey entry, but that the signal for this stage is defective in the Δ*bd1291* mutant, stalling predation at a specific point.

Bd2473 was identified as a homologue of TsaP, a protein which in other bacteria (such as *Neisseria*) stabilises the PilQ porin (through which the pilus extends) in the outer membrane by binding the cell wall peptidoglycan via a lysM domain. The predicted *B. bacteriovorus* protein has significant homology at the N-terminus to TsaP sequences of other bacteria, including the lysM domain, but also has a predicted extended C-terminus which has a predicted transmembrane domain with homology to inner membrane domains, prompting the hypothesis that this protein may span the periplasm and be involved in signalling about the pilus status, a critical point of *B. bacteriovorus* prey-invasion sensing (Capeness et al., 2013). Deletion of this gene could also only be achieved by rescuing HI mutants and was found to be severely hampered in predation, with the vast majority of cells stuck to the point of irreversible attachment to the outer layers of the prey, similar to the phenotype observed with the Δ*bd1291* mutant. Some of the prey were rounded up, suggesting that they had sensed that they were attached to suitable prey cells and had progressed in the predation cycle to the point of secreting the DacB-like carboxypeptidases to effect this (Lerner et al., 2012). A small percentage of the mutant *B. bacteriovorus* had even got as far in the predation cycle as entering the prey, but seem to have stalled in the predation cycle and failed to continue to grow. This suggests that the Δ*bd2473* mutants were defective at signalling at a point similar to that of the Δ*bd1291*, but that some of this signalling was still occurring in a stochastic manner in the population, with most cells stuck on the outside, some prey rounding occurring, and some Δ*bd2473* cells even capable of prey entry.

Thus we had identified a number of mutant strains defective at different stages of pilus-mediated signalling and sought to transcriptionally profile these at the early stages of predation (15 minutes) where this signalling is essential for predation to continue. As all of the mutants necessarily had to be grown as HI mutants and therefore controls for this experiment also had to be HI mutants.

Materials and Methods

RNA preparation of sequencing

RNA preparations for Experiments 1-4 were as described in our earlier work (Lambert et al., 2010a, Capeness et al., 2013). Total RNA was collected and transcription immediately terminated by emersion of the cells in an ethanol/phenol solution which preserved the integrity of the RNA. Extraction was by the Promega SV total RNA kit. Ribosomal RNA depletion, library preparation and Illumina RNA-Seq analysis were performed by Vertis Biolotechnologie AG and the fastq files were provided by them for this analysis.

Computing resources

A Lenovo Thinkstation P500 equipped with an Intel Xeon CPU ES-1620 v3 and 32 Gb RAM was used to run Ubuntu 20.04 and Windows 10 v1809 64-bit.

The Galaxy service, an open source web-based platform for intensive biomedical research, was used at https://usegalaxy.org with a Chrome web browser. 500Gb storage was temporarily assigned to this project upon request.

The iDEP service, an open source web-based platform for implementing R-based analyses was used at http://bioinformatics.sdstate.edu/idep/ with a Chrome web browser (Ge et al., 2018).

Retrieval of fastq files

Files for experiments 1-4 and 8 (supplied by Vertis Biotechnologie AG as described above) were retrieved from our laboratory storage system. Files from Experiments 5-7 were extracted from the Sequence Read Archive (SRA): https://www.ncbi.nlm.nih.gov/sra/ using the fasterq-dump tool on Ubuntu 20.04.

SRA identifiers were renamed (new names to the right of the identifier below) in order to make identification easier for downstream analysis:

Datasets for Experiment 5 (response to *S. aureus* biofilms):

SRR6513822 bd+sa_3

SRR6513821 bd+sa_2

SRR6513820 bd+sa_1

SRR6513819 bd+h_3

SRR6513818 bd+h_2

SRR6513817 bd+h_1

Datasets for Experiment 6 (response to nutrients):

SRR3605970 nb_3

SRR3605969 nb_2

SRR3605968 nb_1

SRR3605967 h_3

SRR3605966 h_2

SRR3605965 h_1

Dataset for Experiment 7 (effects of DSF):

SRR11825140 bd+dsf_2

SRR11825141 bd+dsf_3

SRR11825139 bd+dsf_1

SRR11825136 bd_1

SRR11825137 bd_2

SRR11825138 bd_3

SRR11825135 ap+dsf_3

SRR11825134 ap+dsf_2

SRR11825133 ap+dsf_1

SRR11825132 ap_3

SRR11825131 ap_2

SRR11825130 ap_1

Quality control and Trimming

For quality control and trimming, the fastq files were uploaded to the Galaxy server. Quality control was with the fastQC tool with data from batches of 20-30 fastq files combined with the multiQC tool to pool results.

The trimming tools fastp, Trimmomatic and Trimgalore were tested with default parameters. Trimmomatic was chosen for the final pipeline with the ILLUMINACLIP option and all other options as default.

Trimmed reads were further tested by fastQC and multiQC to confirm quality.

Read alignment and quantitation

The Tool Rockhopper (McClure et al., 2013) was chosen for read alignment and quantitation with default parameters mapping to the *B. bacteriovorus* HD100 genome and with verbose output. This tool is optimised for bacterial RNA-Seq analysis.

Raw read output was converted to comma separated file (.csv) formats and experimental data were merged with the script pandas_merge_inner.py (appendix A). Data were merged using the "Synonym" parameter (which included Bdxxxx numbers for annotated genes). As this parameter was identical for all predicted RNAs, the starting position for RNAs was placed in the Synonym column for these.

For some testing the Rockhopper differential gene expression output was used with $q<1 \times 10^{-5}$. The q value is based on local false discovery rate and choosing this stringent value resulted in a high-confidence dataset.

Data pre-processing, clustering and differential gene expression

The iDEP service was used for pre-processing, clustering and differential gene expression analysis http://bioinformatics.sdstate.edu/idep/ with a Chrome web browser (Ge et al., 2018).

Pre-processing was with the rlog transformation and other parameters as default. Total read counts were analysed for potential depth bias which is automatically flagged in the iDEP output with ratios noted. Scatter plots, bar-and-whisker plots and density plots of transformed data were analysed for data quality.

Hierarchical clustering was carried out with 1000 most variable genes, correlation distance and average linkage with a cut off z-score of 4 and with samples not re-ordered (kept in logical temporal order of the experiment) and with samples normalised by division by standard deviation.

In order to determine the number of clusters for *k*-means clustering, a combination of analysing the hierarchical clustering and the elbow method of plotting the within cluster sum squares versus the number of clusters were used to estimate. Then this number and several surrounding numbers of clusters above and below this number were tested and the number chosen was that which gave good, discrete patterns of expression as would be expected for the experiment. The 2,000 most variable genes were included and genes were normalised by mean centre.

Differential gene expression with the iDEP service was by Deseq2 with a FDR cutoff of 0.1 and a minimum fold change of 2. To test expression patterns of differentially expressed genes in an experiment throughout the predation cycle, the list of genes (and RNAs) were merged with the output of *k*-means clustering from Experiment 1 and then the merge taken for further clustering using the iDEP service.

Figure 2 shows the final pipeline chosen with inputs and outputs.

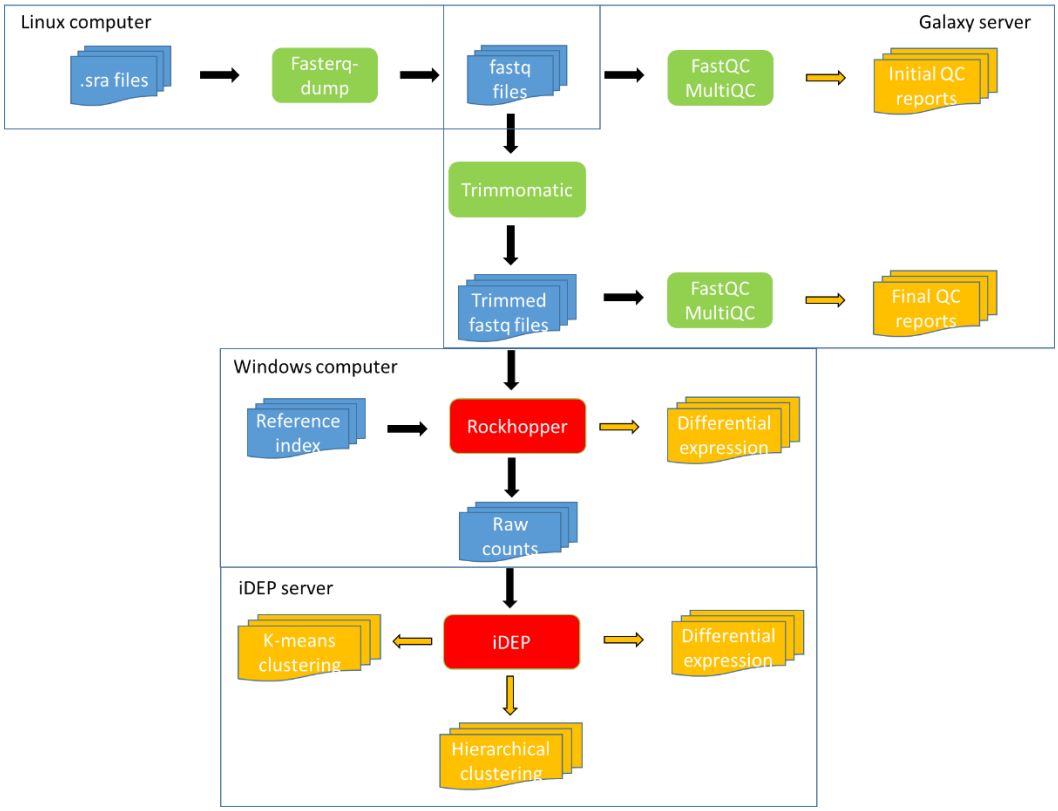

Figure 2- Overview of final analysis pipeline. Outputs of QC reports, differential expression and cluster analysis are shown in yellow. Input files are shown in blue. Individual analysis tools are shown in green. Multi-tool apps are shown in red. Blue boxes indicate the hardware used for the different steps.

Results and Discussion- Pipeline development

Choice of initial experiments to analyse for QC and pipeline development

In order to analyse the data within the computing constraints of the Galaxy server (100 Gb storage and some limitations on processing and memory; although storage was extended to 500 Gb temporarily on request), it was necessary to batch the (>150) fastq files for analysis.

For the first batch, the following experiments were chosen:

Experiment 1 was predation of *B. bacteriovorus* HD100 on *E. coli* K12 (MG1655) throughout the predatory timecourse with samples every 15 min for the first hour, then samples every hour up to the 5 hour timepoint, at which point the prey cells were virtually eradicated and the *B. bacteriovorus* had released into attack phase cells. There were 2 biological repeats at each timepoint for this.

Experiment 5 was comparing *B. bacteriovorus* HD100 in either buffer, or exposed to a biofilm of the Gram-positive non-prey biofilms of *Staphylococcus aureus*. There were 3 biological repeats of these conditions.

Experiment 6 was the third experiment chosen for initial analysis, comparing *B. bacteriovorus* HD100 in either buffer or nutrient broth.

These 3 experiments formed a logical starting point as they are all using the same strain of predator and include the whole predation cycle and different exposures to which the authors had claimed to discover similarities and differences to expression at different points of the predatory cycle (namely that many nutrient utilisation and growth genes were upregulated on exposure to either nutrient broth or biofilms, but that predation-associated genes were not). Further, the 3 different experiments has different sequencing depths and lengths so serve as a good set of examples upon which to test different trimming methods.

Initial quality control

To test the quality of the fastq files, fastQC was run on all files in the first 3 experiments chosen (Experiments 1, 5 and 6) simultaneously, then the results were combined using the tool MultiQC. Table 1 shows an overview of the results. Files for Experiment 5 had the greatest sequencing depth with more and longer reads and have high levels of sequence duplication as a result (highlighted in yellow). K2 samples seemed to have unusually high levels of duplication relative to other samples from this experiment. Files for Experiment 6 have low levels of duplication (highlighted in blue), likely due to the short read length. Figure 3 shows the percentage of overrepresented sequences for each of the samples and this further highlights that the K2 samples are outliers with an excess of overrepresented sequences.

Sample Name	% Dups	% GC	Length	M Seqs
K1-1_fq	68.3%	50%	75 bp	11.2
K1-2_fq	72.2%	50%	75 bp	11.2
K15-2_fq	62.2%	51%	75 bp	9.3
K2-1_fq	79.1%	51%	75 bp	11.8
K2-2_fq	85.2%	52%	75 bp	12.5
K3-1_fq	68.4%	49%	75 bp	13.0
K3-2_fq	69.8%	50%	75 bp	10.9
K30-1_fq	58.6%	50%	75 bp	9.6
K30-2_fq	61.7%	51%	75 bp	10.1
K4-2_fq	65.6%	49%	75 bp	10.9
K45-1_fq	64.1%	51%	75 bp	9.6
K45-2_fq	71.9%	51%	75 bp	10.9
K5-1_fq	70.6%	49%	75 bp	12.0
K5-2_fq	72.3%	49%	75 bp	10.7
KAP-1_fq	72.5%	49%	75 bp	11.3
KAP-2_fq	73.8%	49%	75 bp	10.5
bd_h_1	91.5%	51%	101 bp	26.1
bd_h_2	92.5%	51%	101 bp	26.4
bd_h_3	92.3%	51%	101 bp	26.4
bd_sa_1	89.3%	52%	101 bp	33.8
bd_sa_2	88.0%	52%	101 bp	29.6
bd_sa_3	87.6%	52%	101 bp	28.8
h_1	35.8%	51%	51 bp	20.3
h_2	29.7%	52%	51 bp	17.2
h_3	36.9%	51%	51 bp	21.7
nb_1	25.1%	52%	51 bp	23.3
nb_2	24.3%	52%	51 bp	21.1
nb_3	29.0%	52%	51 bp	20.7

Table 1- Overview of fastq file statistics. K- K12 Experiment 1 AP- attack phase Numbers following K are either time in minutes (15, 30, 45) or hours (1-5) of predation on *E. coli* K12 strain. bd_h and bd_sa; Control (h) and *S. aureus* biofilm exposed samples (sa). h and nb; Control (h) and nutrient broth exposed samples (nb).

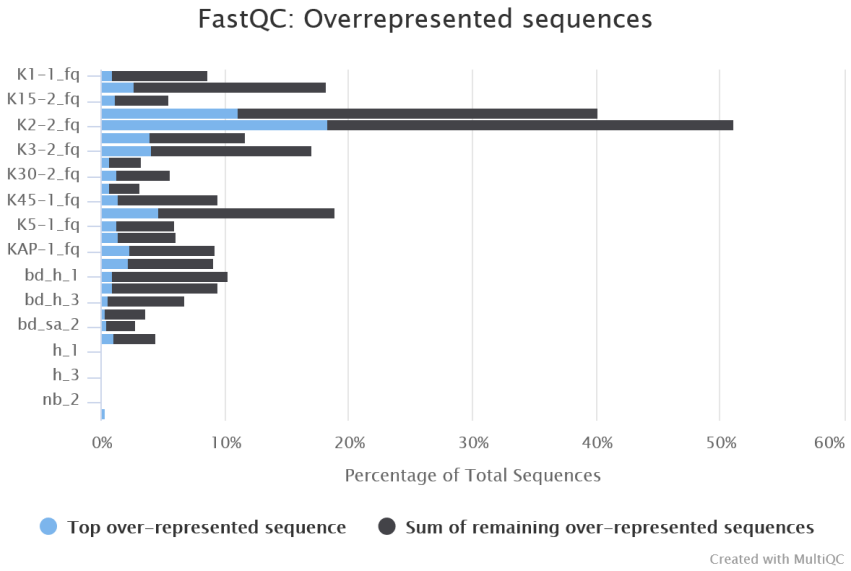

Figure 3- Overrepresented sequences in fastq files.

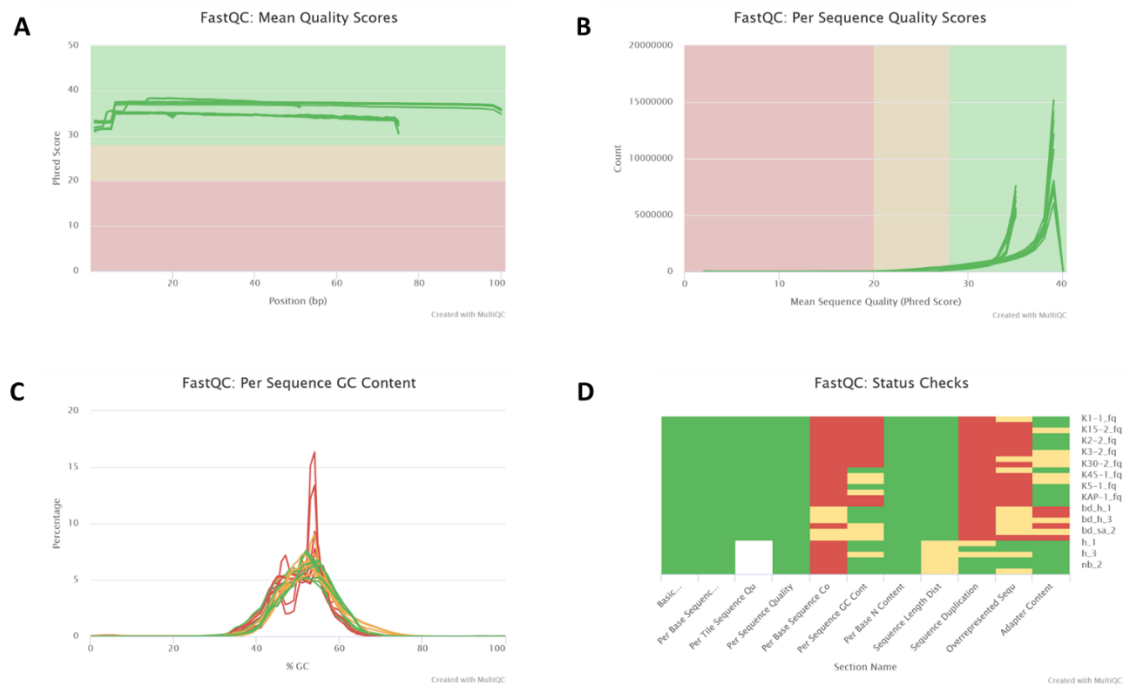

Figure 4- Summary of fastq sequence quality generated by multiQC. A- Mean quality scores across the length of the sequences. **B-** Per sequence quality scores **C-** Per sequence GC Content **D-** Summary of all measured quality parameters

Figure 4 shows that overall, the sequences were of very high quality, with mean quality scores (A) and per sequence quality scores (B) virtually all within good or acceptable range. The different experiments had different levels of quality, with experiments 5 and 6 having overall better quality likely due to greater sequencing depth (having 2-3 times the number of reads compared to Experiment 1). Figure 4C shows the majority of sequences fall within the expected %GC (of ~50%), but that there are some anomalous ones, the most extreme of these are the K2 samples, so these were looked at in more

depth (see below). Figure 4D summarises all of the measured quality parameters. The per base sequence content is flagged in all of the sequences, but this is a common artefact of RNA-Seq due to remaining rRNA, adapter dimers, or biased fragmentation, so is unlikely to be a problem. The other flags of per sequence GC content, sequence duplication and overrepresented sequences are all related and were investigated by checking the K2 data which had the biggest divergence of these.

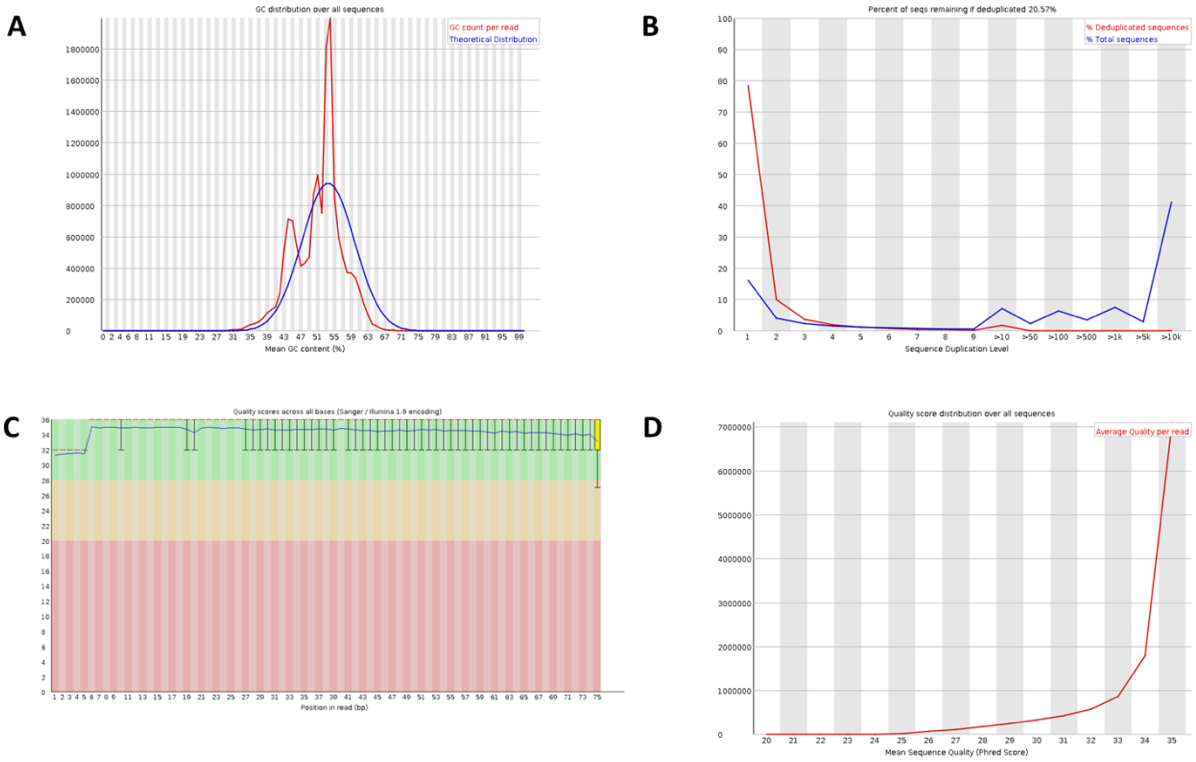

Figure 5- FastQC analysis on K2-1 dataset. A- Mean GC % content **B-** Sequence duplication level **C-** Quality score across all bases **D-** Quality score distribution over all sequences.

Figure 5 shows the fastQC results analysing K2-1 data. This dataset had a large skew in GC distribution from the expected (A) and a large amount of duplication (B). Despite this, the overall quality of the sequence was high (C and D). To investigate the cause of this overrepresentation, the most common sequences were analysed. Figure 6 shows the top represented sequences and their abundance (A) as determined by fastQC. BLAST analysis of the most represented sequence suggest that it is *E. coli* 16S rRNA (B). All of the samples from Experiment 1 were *B. bacteriovorus* preying upon *E. coli* and as there are no means of separating the two experimentally, all samples contain *E. coli* RNA as well as the target *B. bacteriovorus* RNA. The samples were all treated for reduction of both species of rRNA before library preparation by kits using complementary sequence and pull-down techniques, which usually remove >95% rRNA. Throughout the predation cycle, the *E. coli* RNA (both mRNA and rRNA) is degraded by *B. bacteriovorus,* so it is possible that at this 2 hour sample, there are some partially degraded *E. coli* rRNA which was not efficiently removed (as the region targeted for removal may have been degraded). This is a plausible explanation as to why these samples specifically have higher levels of this RNA. Figure 4B indicates that employing deduplication could remove the excessive duplicated sequences, which potentially may improve this sample. However, a study in the utility of deduplication for RNA-Seq found that deduplication improved neither accuracy nor precision and can actually worsen the power and the

False Discovery Rate (FDR) for differential gene expression (Parekh et al., 2016), so deduplication was not pursued for these datasets.

A

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TGTGAAATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGAGGGGGGTAGAAT	1303936	11.041883309340333	No Hit
TTGAAAAATTAGCGGATGACTTGTGGCTGGGGGTGAAAGGCCAATCAAACCGGGAGATAGCTGGTTCTCCCCGAA	448660	3.7993056143619275	No Hit
AAACTGCGAATACCGGAGAATGTTATCACGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAGAGGGAAACA	416732	3.52893555762554	No Hit
CTTAGGCGTGTGACTGCGTACCTTTTGTATAATGGGTCAGCGACTTATATTCTGTAGCAAGGTTAACCGAATAGG	351776	2.978880514861546	No Hit
TTTGAAAAATTAGCGGATGACTTGTGGCTGGGGGTGAAAGGCCAATCAAACCGGGAGATAGCTGGTTCTCCCCGA	189197	1.60214243373698	No Hit
TAGGCGTGTGACTGCGTACCTTTTGTATAATGGGTCAGCGACTTATATTCTGTAGCAAGGTTAACCGAATAGGGG	187732	1.5897366415445848	No Hit

B

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Escherichia coli strain JW5503 chromosome, complete genome	Escherichia coli	139	977	100%	3e-29	100.00%	4630639	CP087136.1
Escherichia marmotae strain Pak2 16S ribosomal RNA gene, partial sequence	Escherichia marmotae	139	139	100%	3e-29	100.00%	1403	ON090427.1
Escherichia coli strain 3_27-F_F09_18 16S ribosomal RNA gene, partial sequence	Escherichia coli	139	139	100%	3e-29	100.00%	859	ON090409.1
Escherichia coli M719 DNA, complete genome	Escherichia coli	139	873	100%	3e-29	100.00%	4856054	AP023433.1
Escherichia coli strain EC13 16S ribosomal RNA gene, partial sequence	Escherichia coli	139	139	100%	3e-29	100.00%	1271	ON077033.1

Figure 6- FastQC analysis of overrepresented sequences. A- top six overrepresented sequences and their abundance. **B-** BLAST hits of the top represented sequence.

Trimming

Pre-processing of fastq files by trimming of adapters and low-quality reads is a common approach for improving RNA-Seq analysis (Williams et al., 2016). Here three common trimming tools were tested to evaluate which improved the datasets for downstream analysis: fastp, Trimmomatic and Trimgalore. Four different datasets were used to test the trimming tools: K1-1, bd+h-1 and h1 to represent datasets with different read lengths and depths and K2-1 as it seemed to have anomalous duplication due to the presence of excess prey rRNA. After trimming with default parameters, fastQC analysis was performed and the results compared to each other and untrimmed fastQC analysis. For all three trimming methods, there was negligible effect on duplication levels, N content, G+C content or per tile quality. All trimming methods increased the spectrum of read length as expected, but all resulted in a very sharp peak with the vast majority of sequences very near their original length. The most significant effect was a slight improvement in quality (Phred score per length). Figure 7 shows quality score across the read length for dataset bd+h-1 untrimmed and after trimming with each of the three different tools. This and the other 3 datasets showed a slightly better performance in terms of read quality by Trimmomatic and so this tool was chosen to pre-process all of the data.

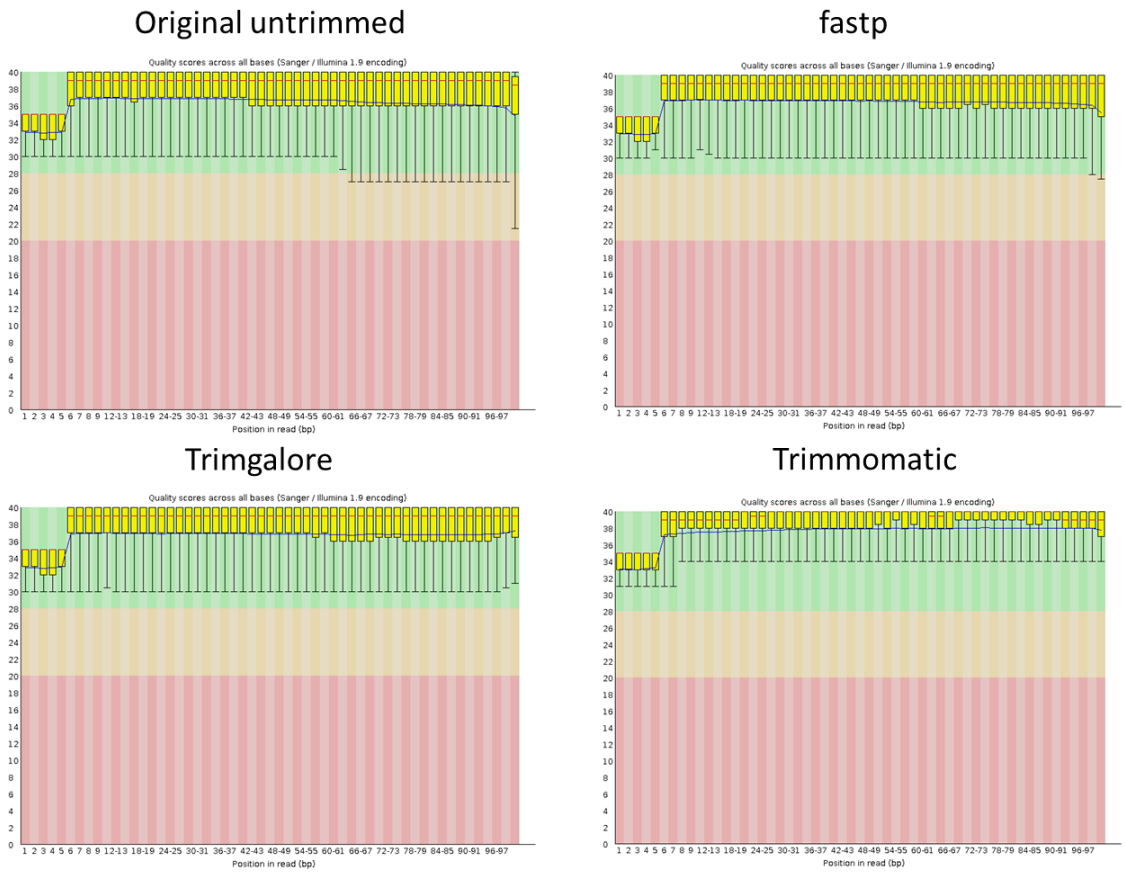

Figure 7- Quality scores per position as determined by fastQC. Quality score (Phred) across the length of the sequence reads is shown for untrimmed reads of the dataset (bd+h-1) and those trimmed with either fastp, Trimgalore, or Trimmomatic with default parameters.

Having chosen Trimmomatic, the use of different parameters within this tool was investigated. Two main non-default parameters were tested: performing an initial ILLUMINACLIP step, which cuts the adapter and other illumina-specific sequences from the read (all datasets were generated by illumina sequencing) and secondly, imposing a quality threshold of Phred 20. The same four representative datasets were chosen as typical datasets to test these parameters on.

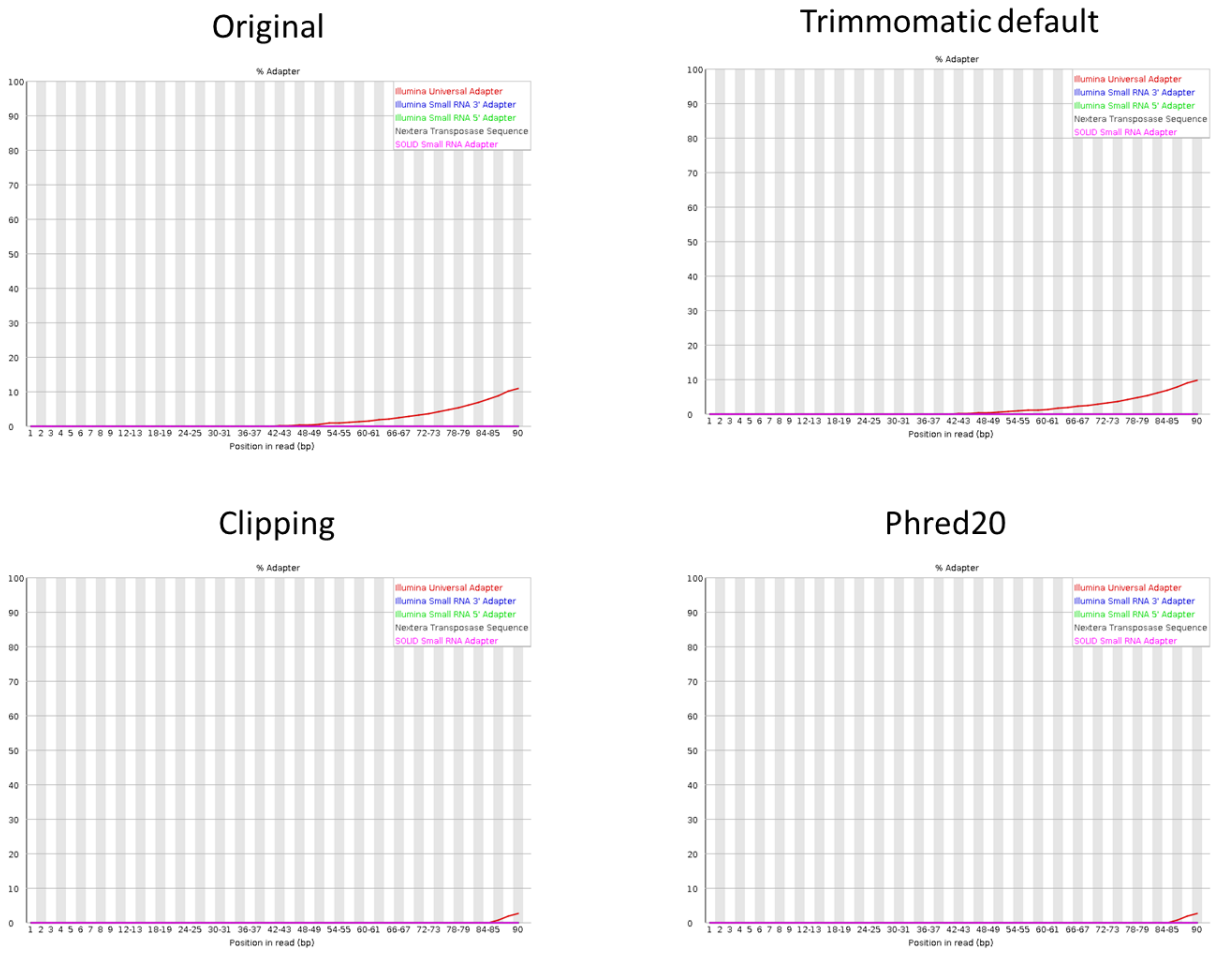

Figure 8- remaining adapter left in reads with different Trimmomatic settings. Results determined by fastQC for dataset bd+h-1 are shown untrimmed, with Trimmomatic default parameters, with ILLUMINCLIP applied or with a quality threshold of Phred20 applied.

Figure 8 shows that both non-default parameters result in greatly reduced adapter in the reads relative to both untrimmed and reads trimmed with default Trimmomatic parameters.

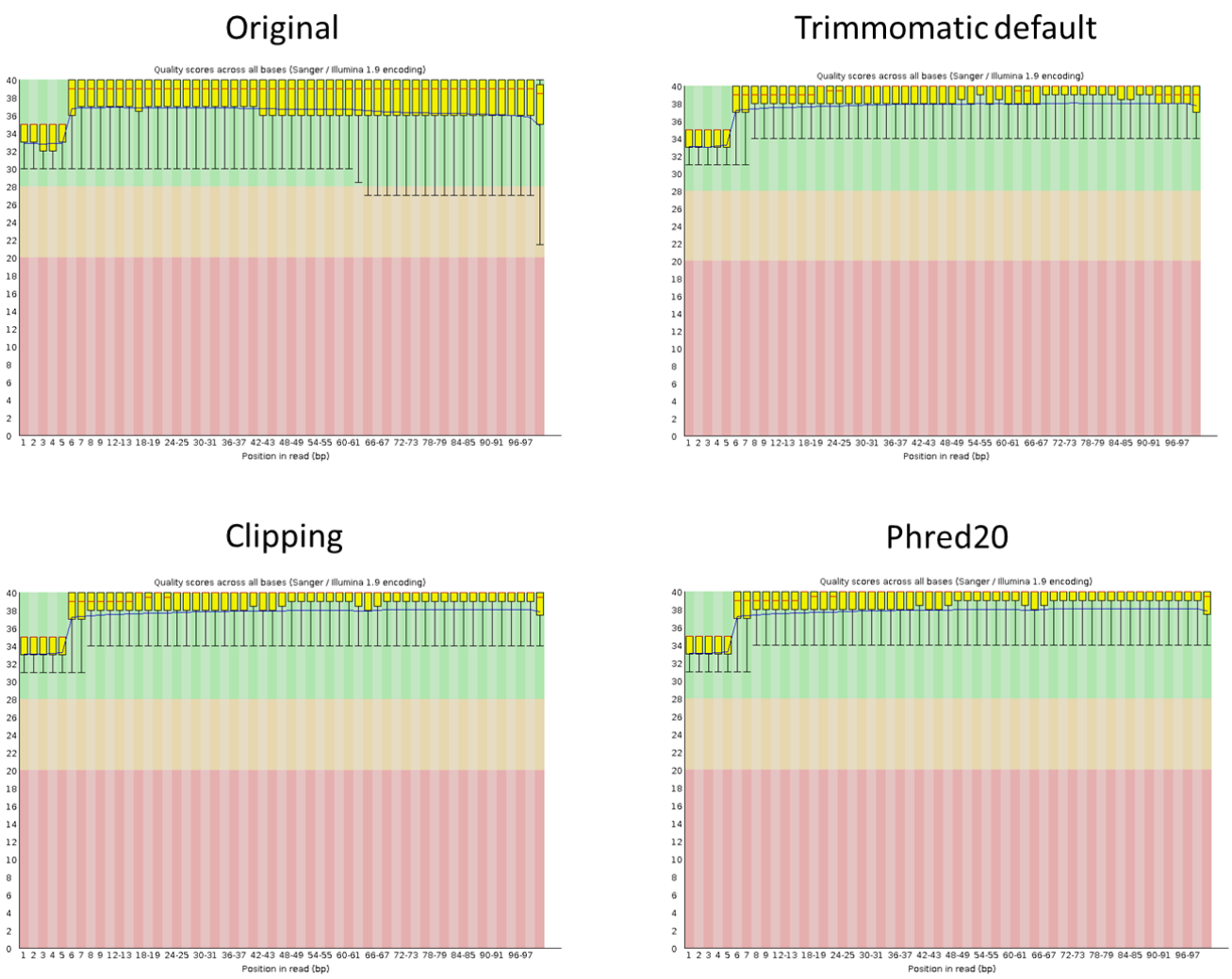

Figure 9- Quality scores per position as determined by fastQC. Quality score (Phred) across the length of the sequence reads is shown for untrimmed reads of the dataset and those trimmed with Trimmomatic with default parameters, or with the ILLUMINACLIP pre-processing or a quality threshold of Phred20.

Figure 9 shows a slight increase in read quality by both non-default parameters. As both methods appear to have a similar outcome, it was decided to proceed just with the pre-processing ILLUMINACLIP step as excessive trimming by quality threshold has been shown to have a detrimental effect on downstream applications (Williams et al., 2016). Therefore, the final pre-processing pipeline was trimming of the reads with Trimmomatic with ILLUMINACLIP and other parameters as default (sliding window trimming), followed by fastQC to check for quality.

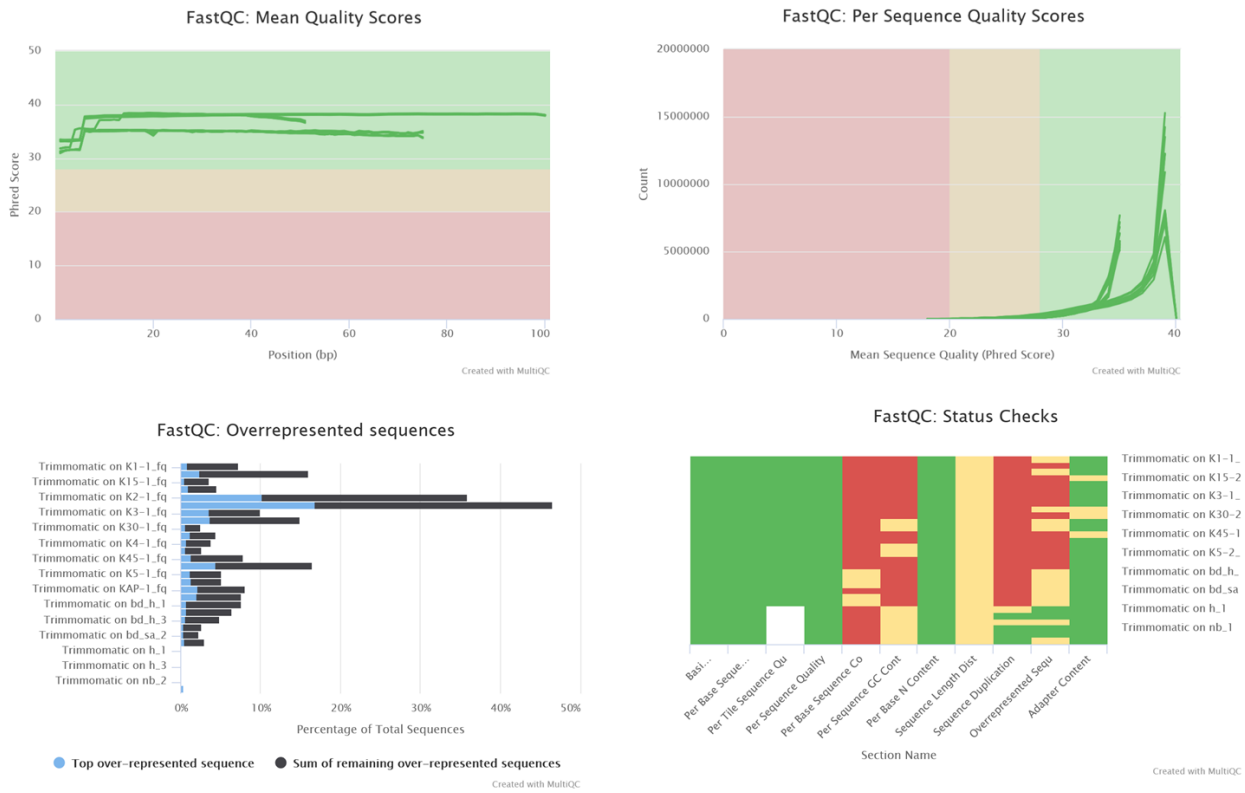

Figure 10- final multiQC of trimmed reads from the first experiments chosen to test. Quality scores, overrepresented sequences and status summary of all trimmed reads from the experiments as output by multiQC.

Figure 10 shows that the final trimmed reads from the first 3 experiments chosen for testing are of high quality, with the only problem being overrepresented sequences which are likely rRNA which should not be detrimental to any downstream analysis. The parameters flagged in the status checks of per base sequence content, sequence duplication and overrepresented sequences are a consequence of this, while per sequence GC count is a common artefact of RNA-Seq and the sequence length distribution is a consequence of different read lengths in the different experiments. In conclusion, the pre-processed data look to be of good quality for further analyses. The data presented are from the first three experiments chosen for analysis, the remaining data were treated to the same trimming and fastQC analysis in batches using MultiQC. These resulted in similar quality reports, with the only flagged parameters the same as described here. Again, analysis of sequence overrepresented sequences suggested that these were unlikely to affect downstream analyses. Also, the overall sequence quality was again very high for these batches.

Sequence mapping and quantification

For alignment of reads to the *B. bacteriovorus* HD100 genome and counting of reads, the tool Rockhopper (McClure et al., 2013) was chosen as this has a pipeline optimised for bacterial RNA-seq analysis. Reads are aligned by Rockhopper using a modified method based on BowTie2. Transcript boundaries are then identified by a method unique to Rockhopper and reads are counted, then normalised by RPKM and an optimised method again designed for Rockhopper, based on normalisation to the upper quartile of expressed genes. Outputs include raw reads, RPKM, Rockhopper normalised reads and an expression value, which incorporates the Rockhopper normalised reads from replicate

inputs into one final value. The iDEP server was chosen for downstream analysis of clustering and differential gene expression testing. To test if Rockhopper normalisation pipelines improved iDEP outputs, the different Rockhopper outputs were tested as input into the iDEP server.

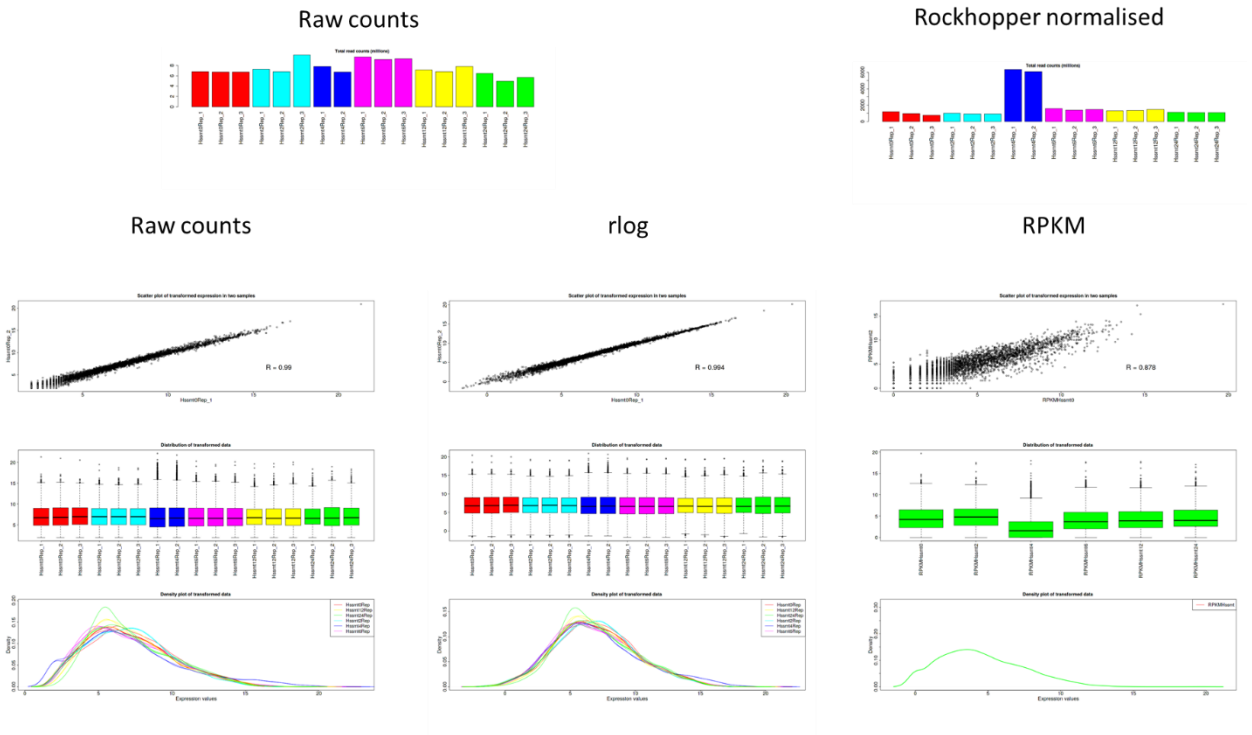

Figure 11- overview of data from Experiment 3. Inputs to the iDEP server were from Rockhopper outputs of normalised data, RPKM or raw counts.

Figure 11 shows that rather than improving the data, Rockhopper normalisation steps appeared to introduce artefacts in the data, with normalised data input resulting in 2 samples (4 hours; discussed later) seemingly having drastically moderated counts. RPKM values lost individual repeat data and also seemed to introduce a skewed distribution with the 4 hour samples (as would the collated Rockhopper "expression" value). Further to this, the iDEP server recommends input of raw counts so this was chosen as the preferred method of input. Figure 9 also shows that the computationally slower transformation of rlog rather than default EdgeR improved transformed data distribution with fewer outliers and improved correlation between repeats, especially for lower expressed genes. The example in Figure 9 shows correlation between reps 1 and 2 for time 0 of Experiment 6 and using rlog improves the correlation R from 0.99 to 0.994, similar improvements were seen for other samples. Therefore the rlog transformation was chosen (with testing for other datasets to confirm good performance with all datasets) as the default pipeline for this analysis.

Results and Discussion- Dataset Analysis

Experiment 1- predation by Wild- Type *B. bacteriovorus* HD100 on *E. coli* K12 MG1655 in buffer throughout the predation cycle.

Experiment 1 was an excess of *B. bacteriovorus* predator mixed with *E. coli* K12 prey to give a single round of semi-synchronous predation and thereby identify the predator genes associated with each stage of prey entry and killing, establishment of the bdelloplast, growth and replication within the bdelloplast, followed by bursting out of the prey and establishing new attack-phase predator cells. The experiment was carried out in Ca/HEPES buffer so virtually all of the nutrients available to the predator are those extracted from the prey.

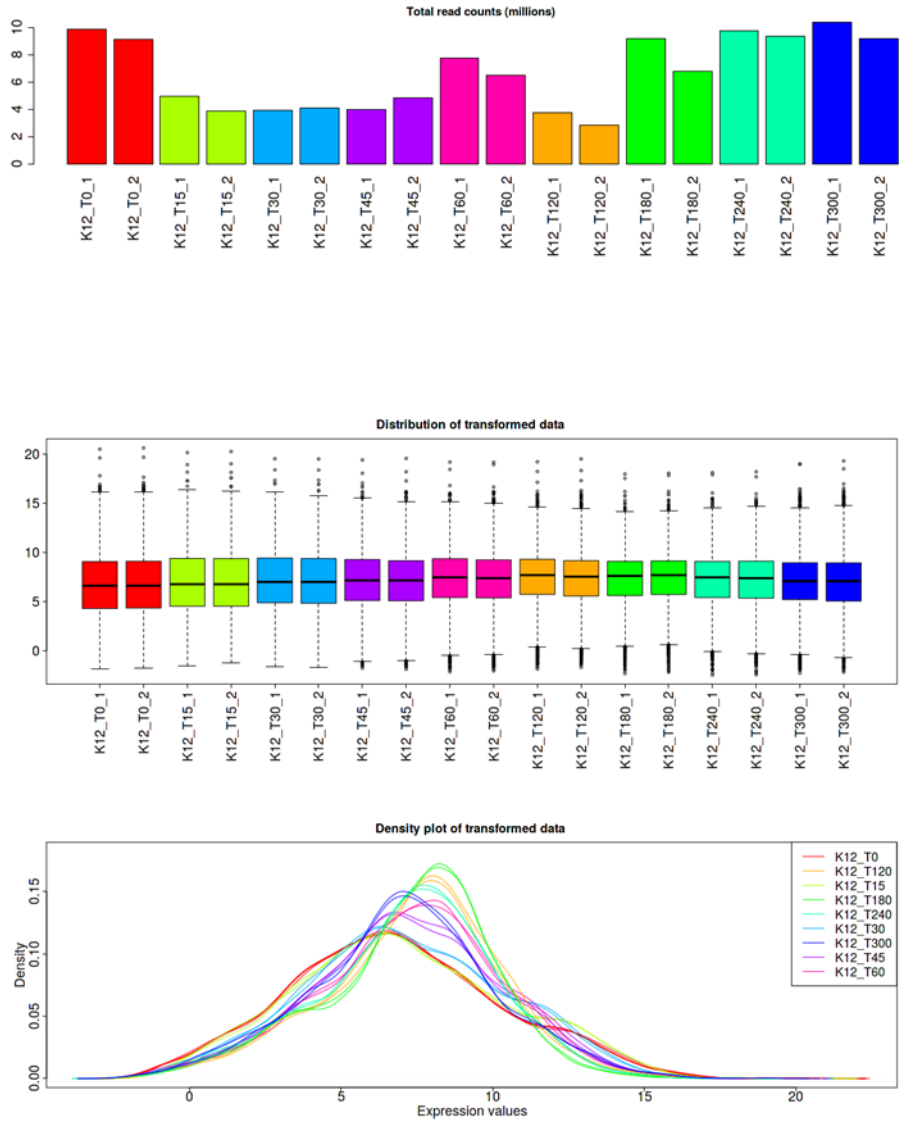

Figure 12- Overview of data for Experiment 1. A- Read counts and distribution of transformed data.

Figure 12 shows an overview of results from Experiment 1. Figure 12 shows good read count numbers (with a minimum of >2M) and normalised (rlog) data distribution. Although it was flagged that there is a potential sequence depth bias detected (with ratios of up to min/max 2.96), this is unlikely to introduce significant bias as the sequencing depth is >25 times for the samples and at least some of the bias is likely a result of differing levels of prey rRNA which will not be mapped. This was true for data from all of the experiments.

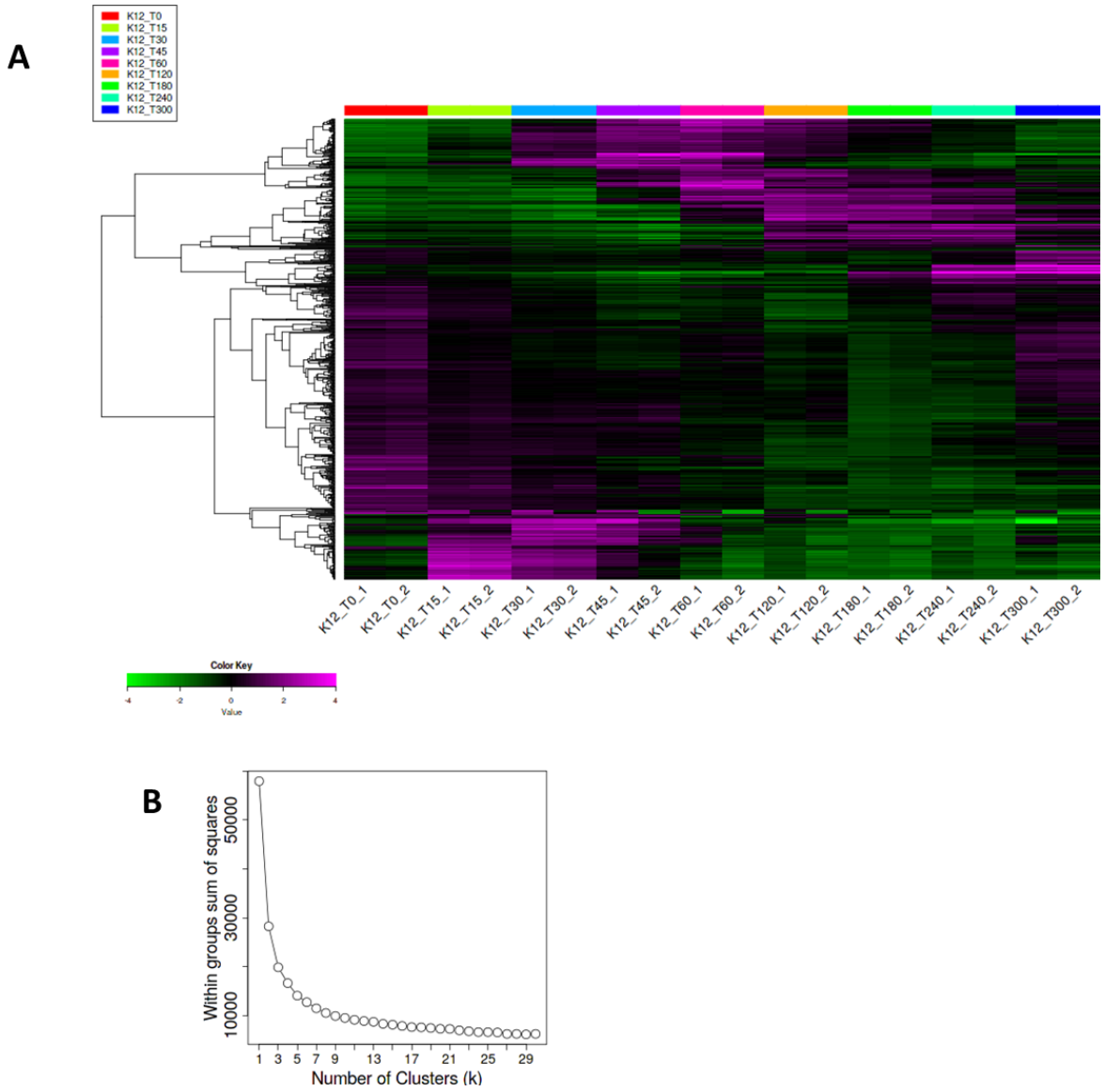

Figure 13 A- Dendrogram of hierarchical clustering with data kept in logical temporal order across the predation timecourse. **B-** Plot of explained difference as a function of clusters to determine optimum number of clusters by the elbow method.

Figure 13 shows a dendrogram of hierarchical clustering of the data from Experiment 1. For this analysis, the data were kept in their logical temporal order of predation throughout the lifecycle. The distance was by correlation method and linkage was average and the samples were normalised. Testing different methods (Euclidean or absolute PCC distance or different linkage methods) did not dramatically alter the results. The methods presented resulted in the most logical-looking layout and so was chosen for these analyses. Our expectation is that the data should arrange into groups of genes differentially expressed at different times during the predation cycle and this

method seems to reflect this very well. Hierarchical clustering broadly arranges the data of Experiment 1 into 3 main clusters: the top group consists of genes whose expression is induced later in the growth cycle, then turned off again, the central cluster is predominantly genes expressed in attack phase, turned off during the predation/growth phase and back on at the end of the cycle where new attack phase cells are generated, and the final cluster consists of genes sharply upregulated early upon prey contact and turned off shortly after. However, there are clearly many subdivisions within these as groups of genes are activated and inactivated at all different stages throughout the cycle. For example, in the top cluster, there are clearly some groups of genes expressed from 15-30 mins, some from 45-60 mins etcetera and it would be informative to distinguish between these likely functionally different groups, which may match with experimental work. Figure 13B shows a plot of explained difference (within groups sum of squares) as a function of clusters and this can be used to determine the optimum number of clusters. Using the elbow method on this plot suggests between 5-9 clusters may be optimal and this along with the hierarchical clustering dendrogram suggest that the higher end of this is likely to be useful in identifying distinct functional groups of differentially expressed genes, so this was borne in mind for *k*-means clustering.

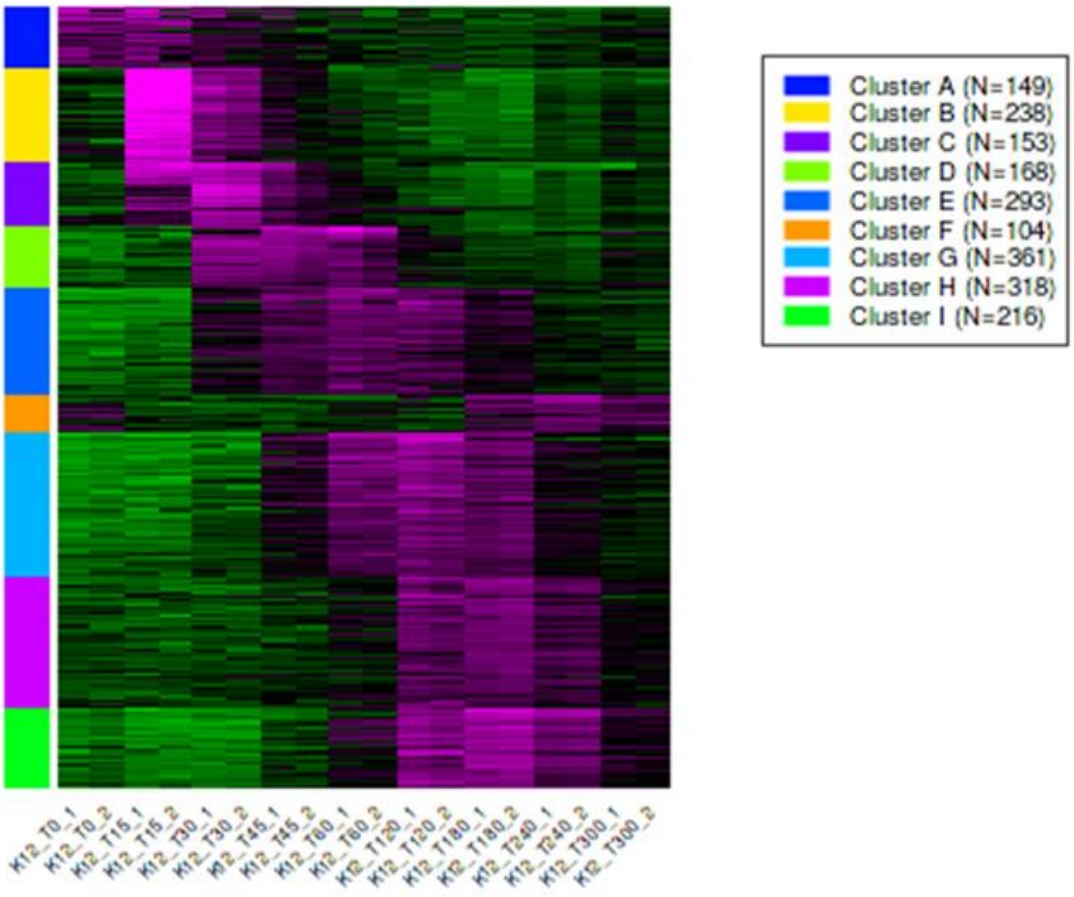

Figure 14- *k*-means cluster analysis of Experiment 1 to 9 clusters

Figure 14 shows the results of *k*-means clustering, choosing 9 clusters and using the default "mean centre" to normalise the genes. Using the "standardisation" method gave similar results whilst the "L1 norm" method introduced anomalies, so the mean centre

method was chosen to use. Each of the clusters represent groups of genes upregulated at specific times throughout the predation cycle as expected.

Cluster D consists of genes upregulated from 30-60 minutes in the predatory cycle. At this point, all of the prey cells are invaded and killed and the predators are beginning to degrade the prey and build their own components for growth. This is reflected in the enrichment for pathways involved in RNA metabolism. The *B. bacteriovorus* degrade the prey RNA (especially the abundant ribosomes), uptake the degradation products and begin to build their own ribosomes at a rapid rate at these timepoints. Also within this cluster is *bd1904*, a gene identified as highly upregulated at 30 minutes by microarray (Lambert et al., 2010a) and whose product is exported into the prey periplasmic space and is hypothesised to form a scaffold-like structure by polymerising. Interestingly, peak expression of this gene by RT-PCR was seen at 15-45 minutes upon predation of the *E. coli* strain S17-1 (which is historically the most common lab strain used in *B. bacteriovorus* research; unpublished observations from our lab). However, this study of predation on *E. coli* K12 shows peak expression of this gene at 45-60 minutes. This highlights the potential differences of predation on even closely related bacteria. This cluster also includes several non-coding RNAs of unknown function, some peptidases and genes associated with iron uptake; the siderophore producing *uicC* and the uptake outer membrane protein encoded by *tonB.*

Similarly, cluster E consists of genes upregulated from 45-180 minutes which represents the second phase of prey breakdown and predatory component building. Here, many pathways for carbohydrate and protein metabolism are enriched which represents the degradation of prey macromolecules and incorporation of the products into predator growth. The cluster analysis has accurately resolved the different phases of first building ribosomes (cluster D), then expression via these of the next phase of predatory digestion and growth enzymes (cluster E). This cluster also contains *dnaA*, the very first step in genome replication.

Cluster G consists of genes upregulated from 60-180 minutes and these are enriched for pathways in energy metabolism, DNA metabolism and organic acids synthesis. These represent the genes involved in predator growth and replication with DNA synthesis and wall and lipid genesis as the predator grows as an elongated filament at these timepoints before fragmenting onto individual cells. This cluster also includes *lamB*, the OMP for uptake of maltose, which had been shown to be upregulated at these times during predation on *E. coli* S17-1 (Lambert et al., 2009), presumably as complex sugars have now been broken down to constituent parts and are beginning to be taken up by the growing *B. bacteriovorus* for use as an energy source. Along with genes associated with the early stages of DNA replication (*dnaX* and *polC*) there are lots of genes in this cluster involved in protection and repair of DNA: *endA, mutS, mutL, recJ, uvrC, ssb, smc, recR, dnaJ, recN, recX* and *recF.*

Finally, Clusters H and I consist of genes upregulated at the end of predation 120-240 minutes and pathways enriched here are for redox and energy metabolism. These may be involved in storing up any remaining nutrients that were not used in replication at the end of the growth cycle and also preparing the newly formed attack phase cells for a change in redox potential as they burst from the exhausted prey cell. Also in these clusters are genes for DNA biosynthesis, indicating that DNA replication is occurring.

As expected, the known pathways of annotated genes are involved in metabolism, growth and replication and so these enriched pathways were all during the growth and replication phases of the predatory cycle from 30-240 minutes. However, there are clusters with genes expressed outwith these timepoints.

Cluster A consists of genes expressed in attack phase that continue to be expressed during the initial interaction with prey at 15 minutes. In this cluster are genes related to flagellum synthesis, reflecting how important flagella motility is in early prey interaction and some chaperones, possibly reflecting different chaperone expression between the attack and growth phases (Lambert et al., 2012). Also in this cluster is the operon *bd2224-bd2229* which encodes a putative transmembrane import/export system, which is downregulated only after prey interaction. Another interesting operon is *bd0798-bd0799,* predicted to encode a catalase and its interacting ankyrin domain protein. This may be protecting the attack phase *B. bacteriovorus* as may the product of *dnaK*, also in this cluster. The global regulators of flagellin motility *fliS* and *bd0881* are also in this cluster and may be driving the continued expression of flagella-related genes until the *B. bacteriovorus* has fully entered the prey.

Cluster B is of genes which are sharply upregulated upon contact with prey and are therefore of great interest as these are genes involved in the early predation process. As relatively little is known of this, few of these are automatically annotated, but this cluster includes many genes that our group has studied. These include many associated with prey peptidoglycan modification such as *bd0993* which encodes a peptidoglycan deacetylase, the L,D-transpeptidases *bd1176, bd1358* and *bd3176*, carboxypeptidases *bd0816* and *bd3459,* lytic transglycosylase *bd3575* and lysozyme *bd1411.*

Cluster C also consists of genes upregulated upon prey contact, but the expression of which peaks slightly later at 30 mins. This cluster also includes few automatically annotated genes and also includes several which we have studied and now know to be involved in prey peptidoglycan modification, involved in sculpting the bdelloplast such as the L,D-transpeptidases *bd3376, bd1358, bd0553, bd0599 bd0886,* lytic transglycosylase *bd3285* and deacetylase *bd3279.* This cluster also contains the operon *bd0412-bd0420* which has homology to gliding motility genes and is upregulated on prey contact. Also, the exonuclease *bd1934*, thought to be exported to the prey cell, suggesting that at this early stage, some degradation of the prey is beginning.

The final cluster without enriched annotated pathways is Cluster F, which has genes highly expressed in attack phase and during the generation of new attack phase cells (240+ mins). Genes in this cluster include those involved in pilus construction and outer membrane proteins, likely reflecting the importance of these structures to the attack phase cell in attaching to prey. This cluster also included *ftsZ*, indicating that cell division is activated at this later stage.

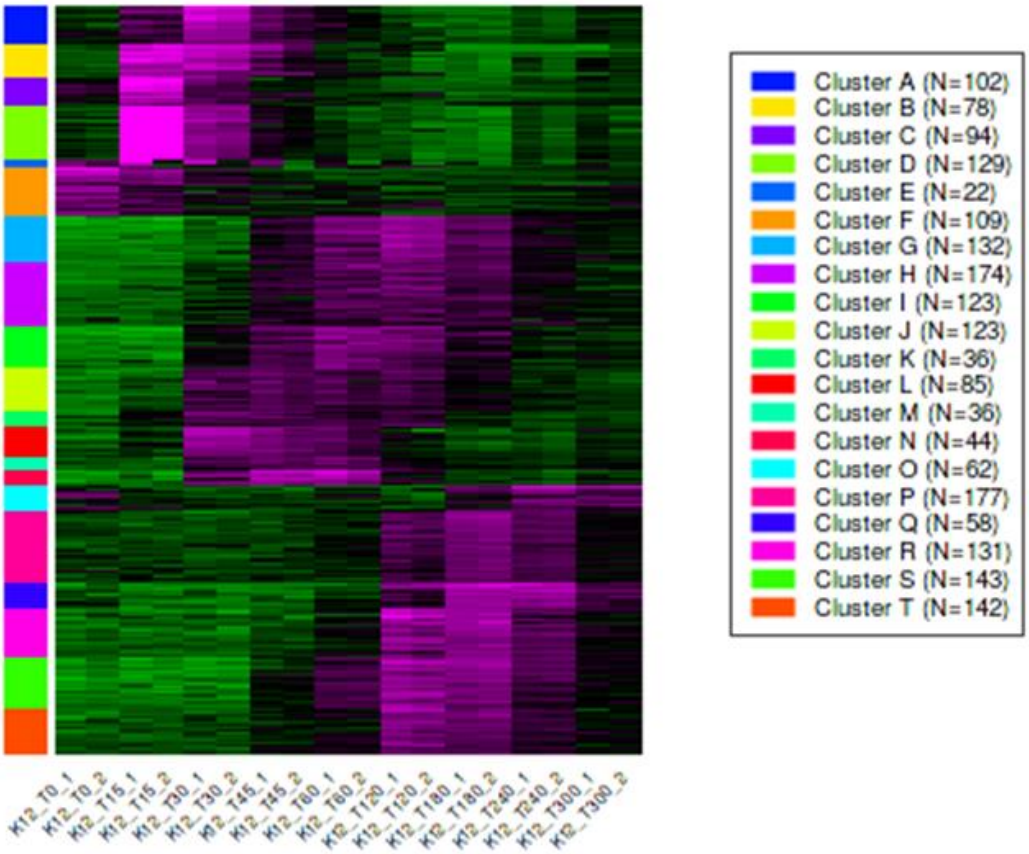

Figure 15- k-means cluster analysis of Experiment 1 to 20 clusters

Increasing cluster number to the maximum of 20 did not improve results (Figure 15), rather it merely split up groups which seemed to have functional cohesion into smaller but similar groups. For example, Cluster G from Figure 14, which consisted of genes involved in energy, DNA and organic acids metabolism, was split into Clusters G and H in Figure 15, with the annotated pathways all falling in Cluster G still. The division of this cluster does not bring any extra information and more clusters result in lower accuracy, so the original choice of clusters based on the hierarchical cluster dendrogram (Figure 13A) and the elbow method on the plot of explained difference per cluster (Figure 13B) seems to have worked well. These methods of estimation were used for each experiment, along with testing different numbers of clusters to determine the optimum for each analysis.

Conclusions from Experiment 1

The data from Experiment 1 was of high quality and was easily organised into functionally sound clusters. The clusters represented genes fairly sharply expressed temporally throughout the predation cycle and the resolution of this was excellent, with various stages of the growth cycle easily resolved, for example, RNA processing and biogenesis genes were in a cluster expressed slightly before those of carbohydrate and protein metabolism. Importantly, there were also stages of predation for which the genes were not extensively annotated and these represent novel genes involved in different stages of predation e.g. early expressed genes in the predation cycle, many of which have been the fruitful objects of study in our lab, particularly the peptidoglycan modifying enzymes found in Clusters B and C. These data will provide a valuable

database for expression of all of the genes (and RNAs) involved in predation with their expression throughout the predation cycle well resolved. Further interrogation of these clusters is a rich resource and could identify interesting new avenues of experimentation to broaden our knowledge of the predation processes.

Some general analyses of patterns of gene expression can also give valuable insights into *B. bacteriovorus* predation. For example genes expressed in attack phase, but with expression that stays high during initial prey interactions, being turned off later after prey entry could well be expressing systems involved in this early interaction. Some of these may also reflect the changing environmental challenges at the attack phase cells enter the prey, for example, the catalase Bd0798 is turned off after prey entry, whilst further redox associated genes are turned on just prior the prey exit, suggesting changing redox challenges at different times in the predation cycle. Also, the co-expression of early DNA replication genes with DNA-binding and protection genes suggests that at this stage, the DNA of the predator is in a vulnerable state, knowledge which may be of relevance in use of *B. bacteriovorus* as a therapy. Predation by *B. bacteriovorus* is clearly driven by specific rounds of expression of genes, with the clusters tightly regulated, being turned on at the specific times that their products are needed and off again, when used.

Experiment 2- predation by Wild-Type *B. bacteriovorus* HD100 on *Serratia marcescens* in buffer throughout the predation cycle.

This experiment is similar to Experiment 1, except on a different prey species, *Serratia marcescens,* and with less resolution, having timepoints only at 0, 2, 4, 6 and 12 hours. The intention was to see if predation on a clinically-relevant, multi-drug resistant pathogen would proceed in a manner similar to that on the lab strain *E. coli* in order to examine the possibility of use of *B. bacteriovorus* as a novel antimicrobial therapy.

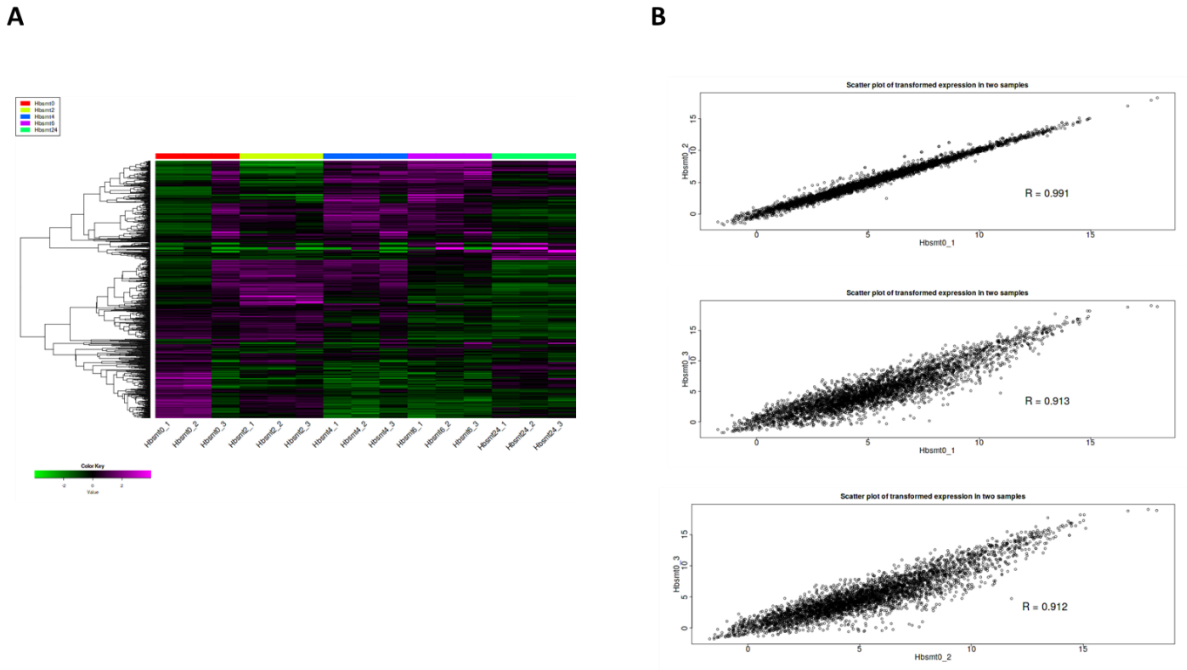

Figure 16A- Dendrogram of hierarchical clustering of experiment 2. **B-** Correlation analyses of time 0 repeats.

Figure 16A shows a dendrogram of hierarchical clustering of Experiment 2, using the same conditions as for Experiment 1. This immediately highlights a problem with the data in that the 3rd repeat of each timepoint seems to be drastically different from the other 2 repeats. Figure 16B shows correlation relations between the 3 repeats at time 0 for this experiment and confirms that repeat 3 is an extreme outlier. Similar results were obtained by analysing the other timepoints (although t=0 shown in Figure 16B was the most extreme example). This is likely a result of the fact that the first 2 repeats were carried out simultaneously, but the third was carried out years later by a different worker. Whilst it is disappointing that a third repeat was not better aligned to the others, there are many difficult biological aspects to this experiment with 2 different organisms and therefore a relatively small difference in any of the culturing and handling of the bacteria could end up with significantly different results. It was therefore decided to exclude the third repeats of this experiment as the most cursory of looks at the dendrogram show that this will clearly obfuscate the results. The analysis was therefore repeated, using only reps 1 and 2.

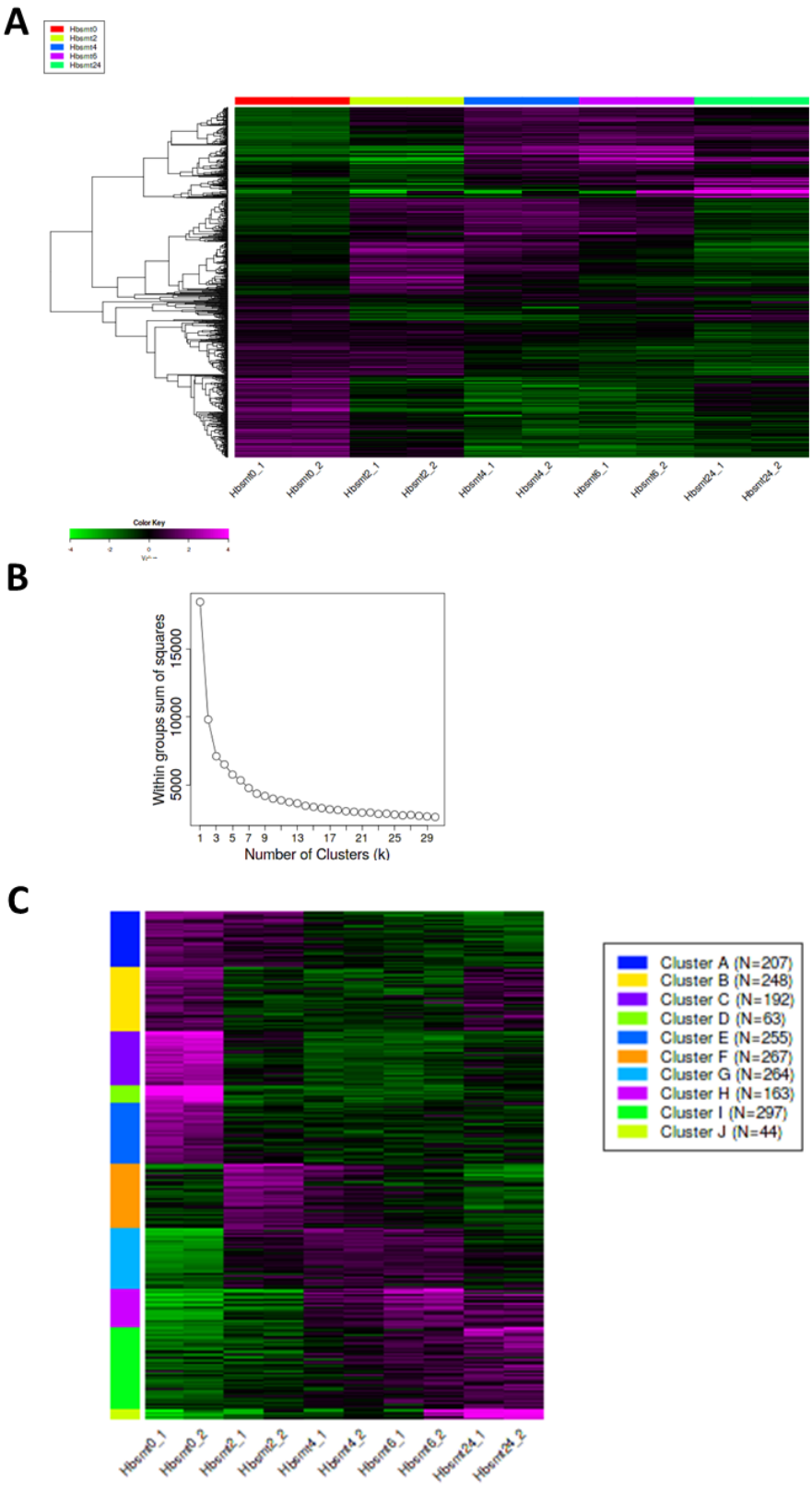

Figure 17 Analysis of Experiment 2. A- Dendrogram of hierarchical clustering with data kept in logical temporal order across the predation timecourse. **B-** Plot of explained difference as a function of clusters to determine optimum number of clusters. **C-** *k*-means clustering into 10 clusters.

Figure 17A suggests that the data from Experiment 2 clusters into slightly less organised groups than that from Experiment 1. This may partly be a result of the lower resolution of the experiment as genes tend to be upregulated in blocks of 2 hour samples as is to be expected. Figure 17B also suggests 7-12 clusters would be appropriate so 10 clusters were chosen for *k*-means clustering, the results of which are presented in Figure 17C.

Cluster B represents genes upregulated during attack phase, turned off during the predation timepoints (2-6 hours) and then on again at 24 hours when attack phase cells predominate again. Puzzlingly, pathways enriched in this cluster include those for transcription, translation and protein metabolism, those which were strongly associated with the growth phase when *B. bacteriovorus* was preying upon *E. coli* (cluster E in Figure 11A for Experiment 1 above, representing genes upregulated at 45-180 mins).

Cluster A represents genes highly expressed in attack phase that stay on upon contact with prey cells (at 2 hours) and this cluster was enriched for flagellin–related genes and chaperones which matched the equivalent cluster when preying upon *E. coli* (Cluster A in Figure 14), albeit that in Experiment 1, these genes were downregulated after 30 mins, but here in Experiment 2, they are still upregulated at the 2 hour timepoint.

Clusters C, D and E are a large group of genes upregulated at attack phase in varying amounts (hence arranging into different clusters), but then downregulated at all other timepoints. There are few annotated genes within these clusters and no obvious themes, so these are intriguing groups for further analysis.

Cluster F are genes upregulated at 2 hours, with some staying upregulated at 4 hours. This cluster is enriched in chemotaxis and motility genes. These genes are again expressed at a different time compared with predation on *E. coli*, where these genes were in Cluster A in Figure 14, expressed at attack phase and into 15 mins after interaction with prey, but then downregulated.

Clusters G and H are genes upregulated from 4-6 hours and again, these seem to be very different from the equivalent timepoints from Experiment 1; there are few growth and division genes expressed here, and indeed some genes associated with attack phase from Experiment 1 are upregulated in these clusters.

Clusters I and J represent genes expressed highly at 24 hour where attack phase cells are again predominant. There are few annotated genes in these groups, but again, some attack phase genes are represented in these groups.

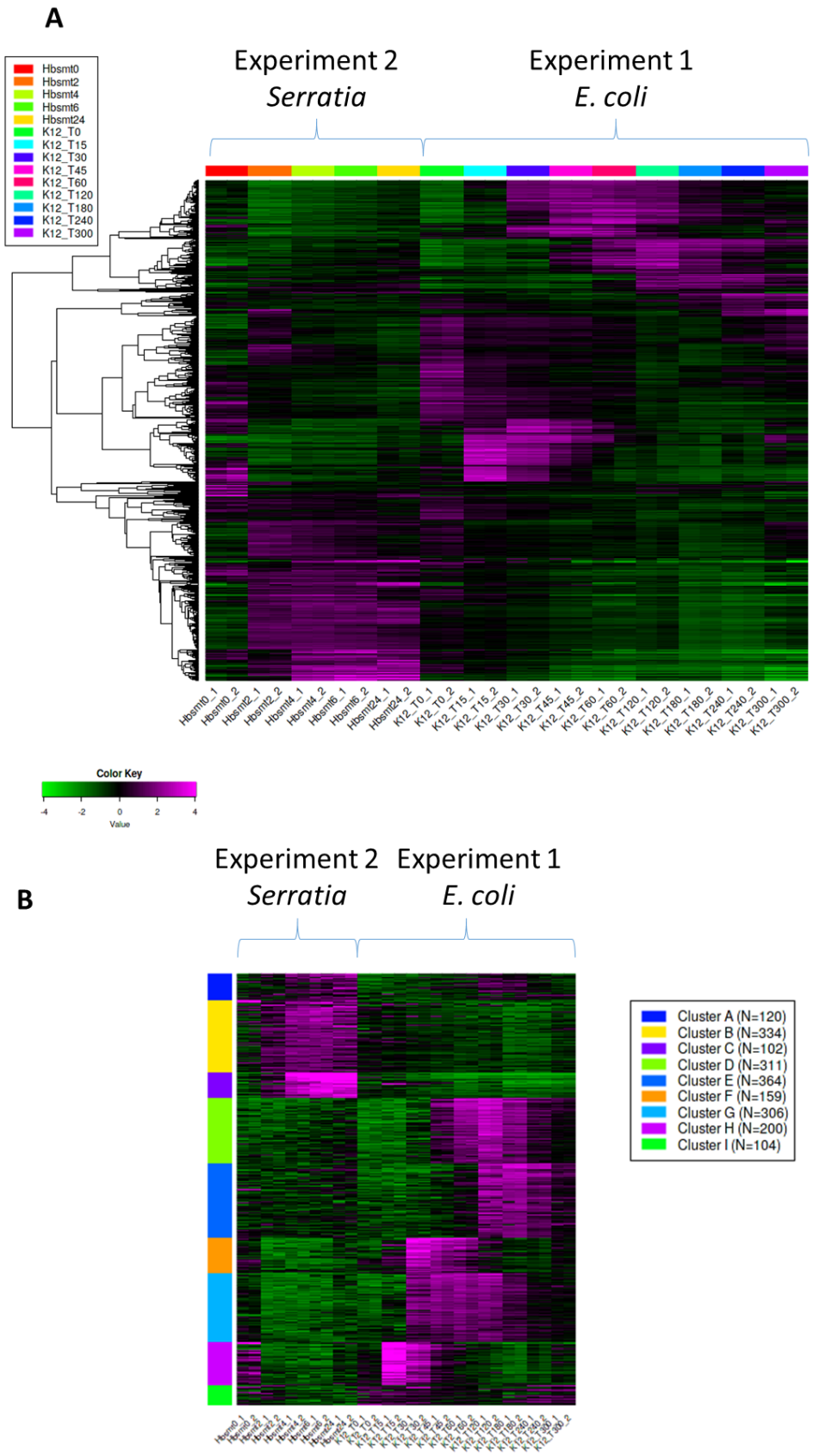

Figure 18- Combined analyses on Experiments 1 and 2 A- Dendrogram of hierarchical clustering B- *k*-means clustering to 9 clusters

In order to further investigate the apparent disparity between gene expression patterns in Experiments 1 and 2, the data were merged and analysed together. Figure 18 shows that there seems to be virtually no overlap of gene expression between the two experiments, with the majority of clusters created based on expression from Experiment 1 and these genes apparently not expressed at any specific point in Experiment 2, likely expressed sporadically throughout the timecourse. Three clusters (A-C in Figure 18B) forming from genes expressed in Experiment 2 were predominantly not expressed in Experiment 1. The partial exception to this is Cluster A which has genes upregulated at later timepoints in Experiment 2 which are also expressed to some extent in later growth phase of Experiment 1, however, even this did not match well as these genes were highest expressed at 24h in Experiment 2 and at 120-180 mins in Experiment 1. Also, there were few annotated genes in Cluster A and no obvious theme to the genes in this cluster. Similarly, there were few annotated genes in clusters A-C and no obvious patterns of gene expression in these clusters which were upregulated throughout the later stages of Experiment 2, but not in Experiment 1.

To further examine this, sets of differentially expressed genes were determined for Experiment 2 and interrogated for their expression patterns in Experiment 1.

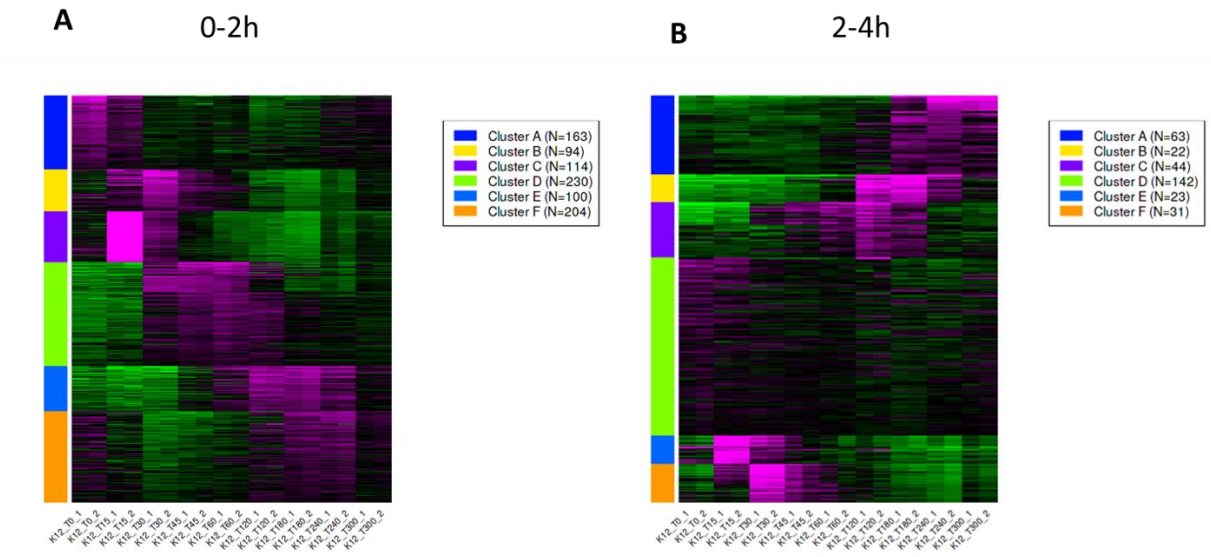

Figure 19- *k*-means clustering to 6 clusters of differentially expressed genes from Experiment 2 mapped onto Experiment 1 **A-** genes from Experiment 2 that were differentially expressed from 0-2 hour timepoints, **B-** genes from Experiment 2 that were differentially expressed from 2-4 hour timepoints.

Figure 19 shows *k*-means clustering of gene expression in Experiment 1, for the subset of genes that are significantly differentially regulated between 0-2 hours (A) and 2-4 hours (B) in Experiment 2. If predation in Experiment 2 was occurring synchronously, at a rate similar to that of Experiment 1, then these genesets would cluster to specific times during the timecourse. In both cases, clusters of gene expression of these datasets are across the whole of the timecourse in Experiment 1, suggesting that in Experiment 2, predation is occurring asynchronously, relative to Experiment 1, with sets of genes at all points throughout the timecourse in Experiment 1 being expressed at each 2 hour timepoint in Experiment 2. This renders it impossible to accurately compare expression between the two experiments to determine small differences between predation on the two different prey types.

Conclusions from Experiment 2

The surprising lack of correlation between genes expressed throughout the predation cycle of Experiment 1 with those across the timecourse of Experiment 2 suggests a difference in the predation on the lab strain *E. coli* (Experiment 1) and the multidrug-resistant, clinical isolate of *Serratia marcescens* (Experiment 2). As there were no cohesive groups of functional genes associated with specific growth or reproduction phases obviously upregulated at specific points in the interaction with *Serratia*, this suggests that the predation was not occurring synchronously as it was on *E. coli*. Rather, it seems that predation was asynchronous with different predators at different stages of growth at the same timepoints, and hence groups of genes were not strongly upregulated at specific points from population samples. The *B. bacteriovorus* had previously been grown on *E. coli* before both experiments and thus was likely ready for synchronous infection of further *E. coli*, but this may not have happened for predation on *Serratia*. In addition to the switch in prey, there were some differences in initial predator concentrations between experiments and also, the smaller *Serratia* were consumed at a faster rate than *E. coli*, with most prey lysed by 2-3 hours, hence at the 2 hour timepoint, genes associated with later timepoints in Experiment 1 were upregulated. These are important conclusions to observe that predation on different, clinically relevant strains is different from that of the lab paradigm, as if *B. bacteriovorus* were to be used as a biocontrol against such organisms, the way in which they behave in these circumstances needs to be understood in detail. Future improvements on Experiment 2 may be to pre-grow the *B. bacteriovorus* on *Serratia* prey to acclimatise them in order to aid synchrony of the predation for measurement of transcription and for more timepoints earlier on in the lifecycle.

Experiment 3- predation by Wild-Type *B. bacteriovorus* HD100 on *Serratia marcescens* in pooled human serum samples throughout the predation cycle.

This experiment is similar to Experiment 2, with predation on *Serratia marcescens,* with timepoints at 0, 2, 4, 6 and 12 hours, only instead of buffer, the medium was pooled samples of human serum. The intention was to see if predation in a medium more aligned to clinical conditions had any effect on predation compared to the more ideal lab conditions of buffer.

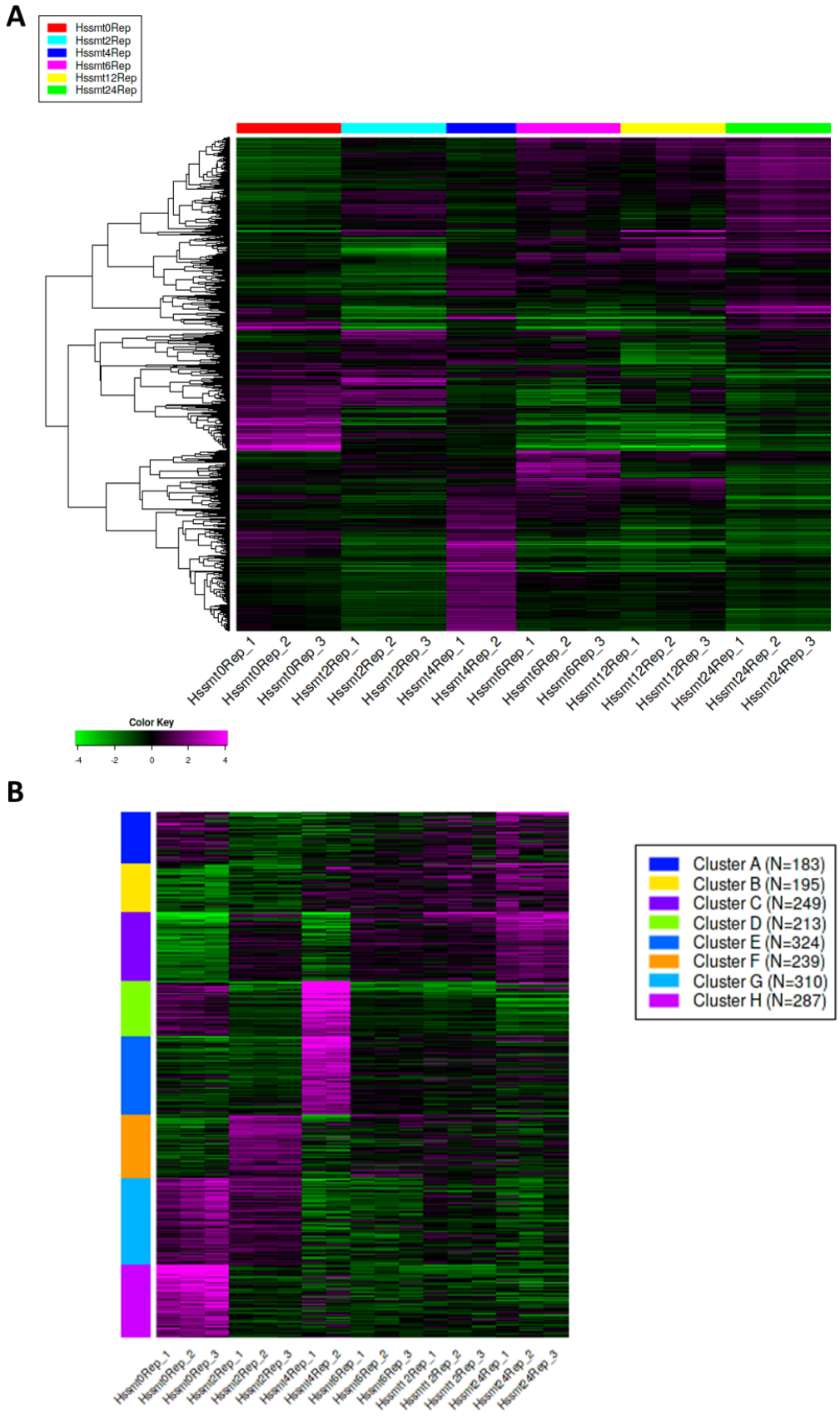

Figure 20 A- dendrogram of hierarchical clustering for Experiment 3 **B-** *k*-means clustering to 8 clusters for Experiment 3.

Figure 20 shows that clustering for Experiment 3 was similar to that of Experiment 2, with blocks of genes expressed predominantly in one timepoint. The only pathways enriched for were those of general transcription, growth and metabolism and again similar to Experiment 2, these were in Cluster G, which are upregulated at 0 and 2 hours, which again is in contrast to Experiment 1. Again, similar to Experiment 2, genes in other clusters did not form obvious functional groups. It is interesting to note a significant difference in expression at 4 hours compared to the rest of the timepoints. It is thought that at this point in the experiment, many antimicrobial compounds (such as complement) have been used up, being absorbed by the bacteria or degraded and at this point, the *Serratia* begin to grow back in what has now become a rich growth medium. The *B. bacteriovorus* may then respond to the differences in prey recovering and growing and indeed, in this cluster are some genes for global transcriptional regulation, including the alternative sigma factor rpoE and several DNA-binding proteins.

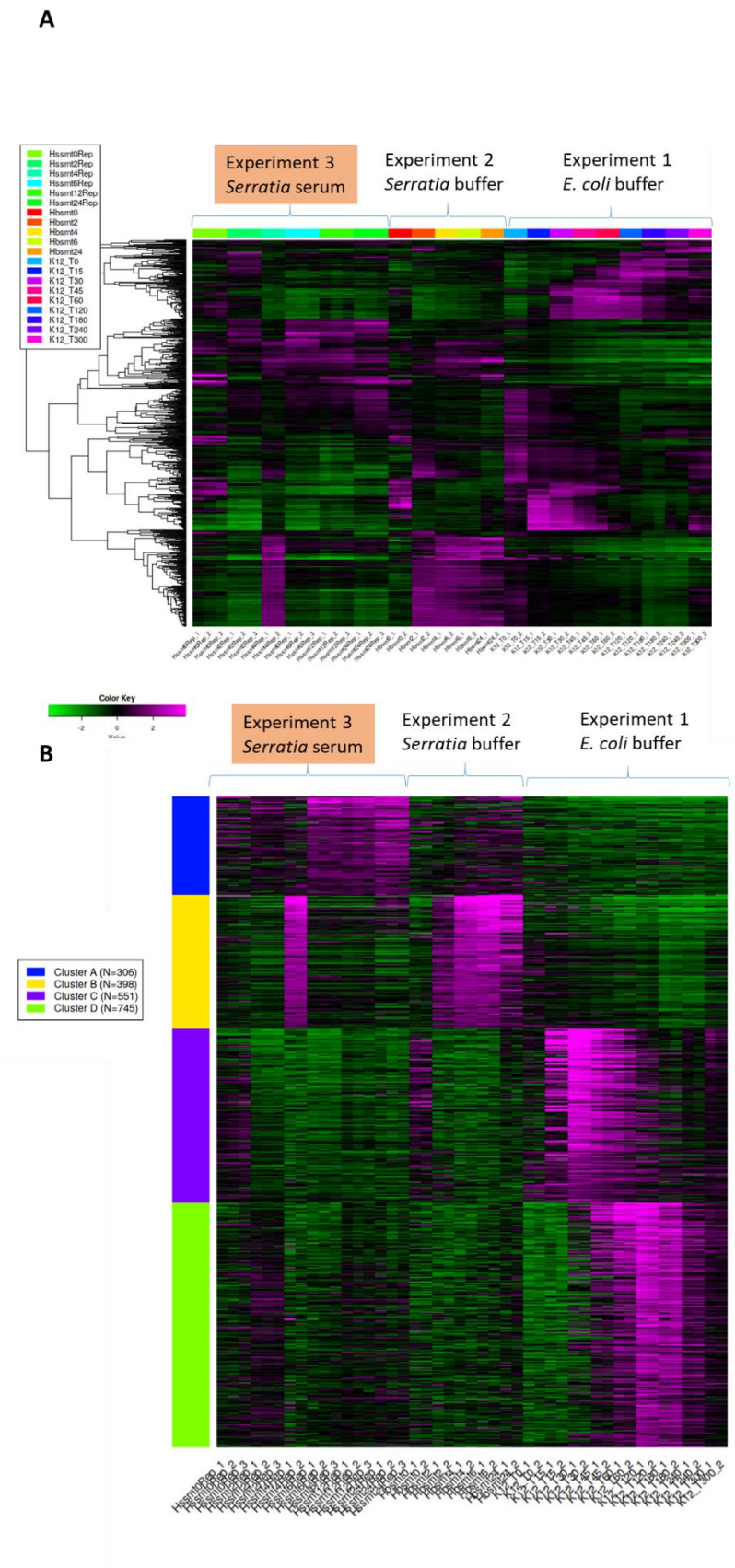

Figure 21 A- dendrogram of hierarchical clustering for Experiments 1-3 **B-** *k*-means clustering to 4 clusters for Experiments 1-3.

To further examine the expression patterns, data from Experiments 1-3 were merged and analysed. Figure 21 further highlights the unexpected stark differences in expression between the 3 experiments. Again, of particular note is that the genes of known function expressed throughout the growth phase for predation on *E. coli* in Experiment 1 are scarcely synchronously upregulated at any point in Experiments 2 and 3. Cluster C is enriched for pathways of gene expression, protein and general metabolism, which are upregulated from 15-120 mins in Experiment 1. Figure 21B shows that a proportion of these, but not most, are upregulated at time 0 in both Experiments 2 and 3, suggesting that these are housekeeping genes expressed during attack phase, but highly upregulated during the synchronous growth phase in Experiment 1. Similarly, Cluster D is enriched for genes involved in energy metabolism and ion transport which are upregulated later (120-240 mins) in Experiment 1 and subsets of this cluster are upregulated at several different timepoints (mostly 2 hours and 12-24 hours) in Experiment 3.

Interestingly, there is a large cluster of genes in Cluster B which seem to be expressed only in Experiments 2 and 3 (preying upon *Serratia*) and predominantly not in experiment 1 (preying upon *E. coli*). These are expressed from 2-12 hours in Experiment 2 (in buffer), but only at 4 hours in Experiment 3 (in serum, at the timepoint where the serum changes composition). Also, Cluster A consists of genes predominantly expressed in Experiment 3, with some expression in Experiment 2 and virtually no expression in Experiment 1. Figure 21B gives a good overview of the drastic differences in gene expression between the 3 experiments, in order to examine this in greater detail, more clusters were generated.

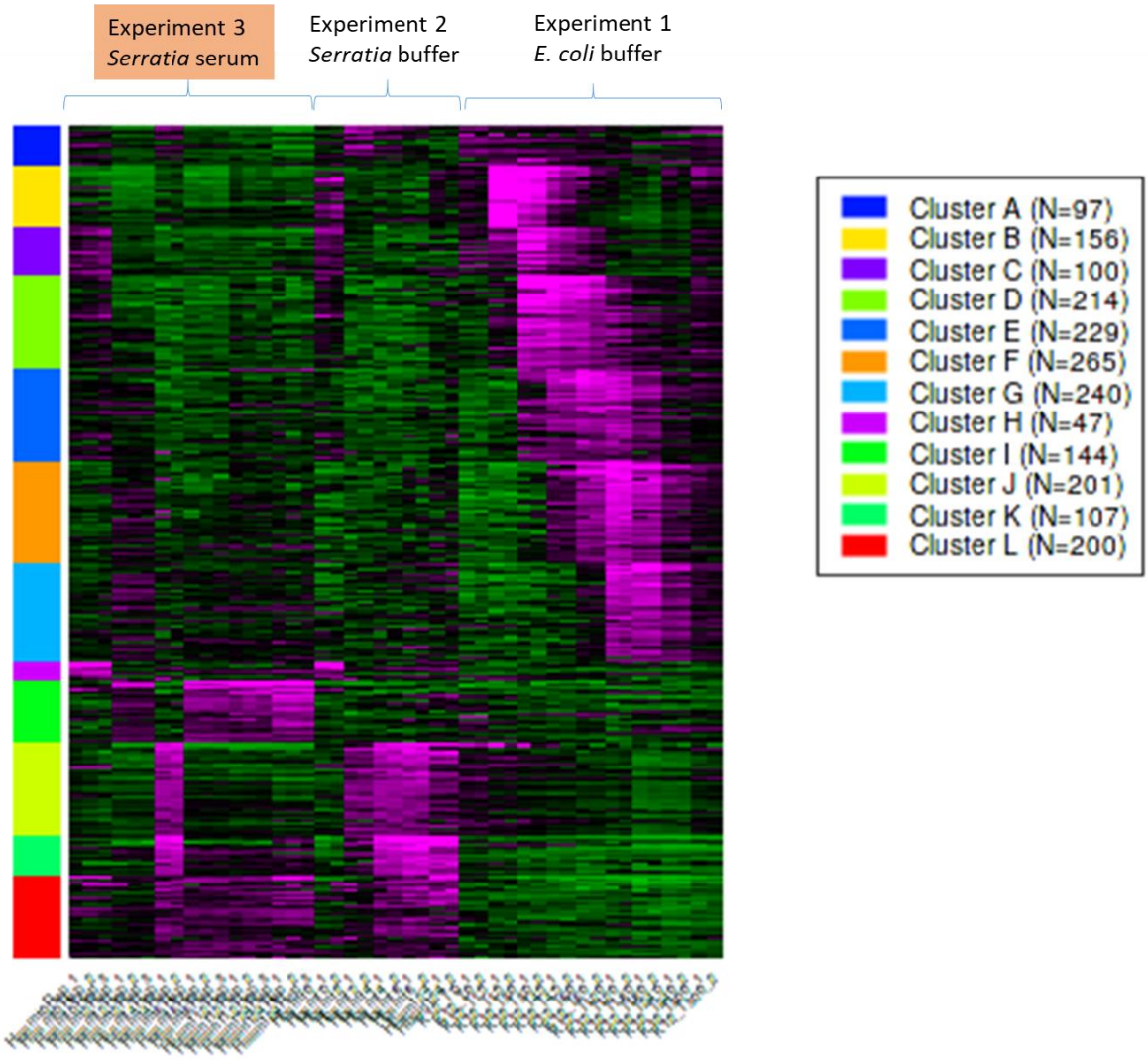

Figure 22- *k*-means clustering to 12 clusters for Experiments 1-3.

Figure 22 shows the results of clustering the data from all 3 experiments into 12 clusters. Cluster A consists of genes upregulated throughout the early and late stages of predation in Experiment 1 and these are upregulated at different, discrete times in Experiments 2 and 3. This group includes flagellin and chemotaxis genes, so this group may represent expression by those excess of *B. bacteriovorus* that are still in attack phase (an excess of *B. bacteriovorus* was added in each experiment at a ratio of 2-4:1 predator:prey in order to attempt to achieve semi-synchronous predation; this was achieved in Experiment 1).

Clusters B-G consist of genes upregulated at specific points from early to late in the predation process in Experiment 1 and Figure 22 highlights again that these are not forming distinct upregulated groups at similar times in experiments 2 and 3.

Cluster H is a small set of genes upregulated only at time 0 upon mixing with prey in Experiments 2 and 3 and not in Experiment 1. These likely represent housekeeping attack phase genes rather than a response to this mixing as the RNA was harvested immediately before any response could be elicited. As these housekeeping genes are highly upregulated in later stages in Experiment 1, normalising resulted in these being at

an apparently low level in attack phase, although these are likely also expressed in attack phase in Experiment 1.

Of particular interest is Cluster I, which consists of genes upregulated only in human serum (Experiment 3) and not in buffer (Experiments 1 and 2). Amongst these are genes involved in iron metabolism, including siderophores (likely responding to the high iron levels in serum) along with permeability related genes such as outer membrane proteins and ion transporters, again likely responding to the challenges of the much denser environment which also contains antimicrobial agents.

Finally, Clusters J-L consist of genes upregulated throughout the timecourses of Experiments 2 and 3, but scarcely at all in Experiment 1 (except some from Cluster J are expressed in attack phase), so these may represent genes involved in interaction and predation upon *Serratia*, but not on *E. coli*. In amongst the many genes of unknown function are many proteases and chaperones, possibly suggesting that *B. bacteriovorus* may use different sets of analogous genes for predation upon different prey. The genes from Clusters J-L which are also expressed in attack phase in Experiment 1 include many genes for flagellar motility and pili generation (including *pilA* and many minor pilins) and the operon *bd2224-bd2229.*

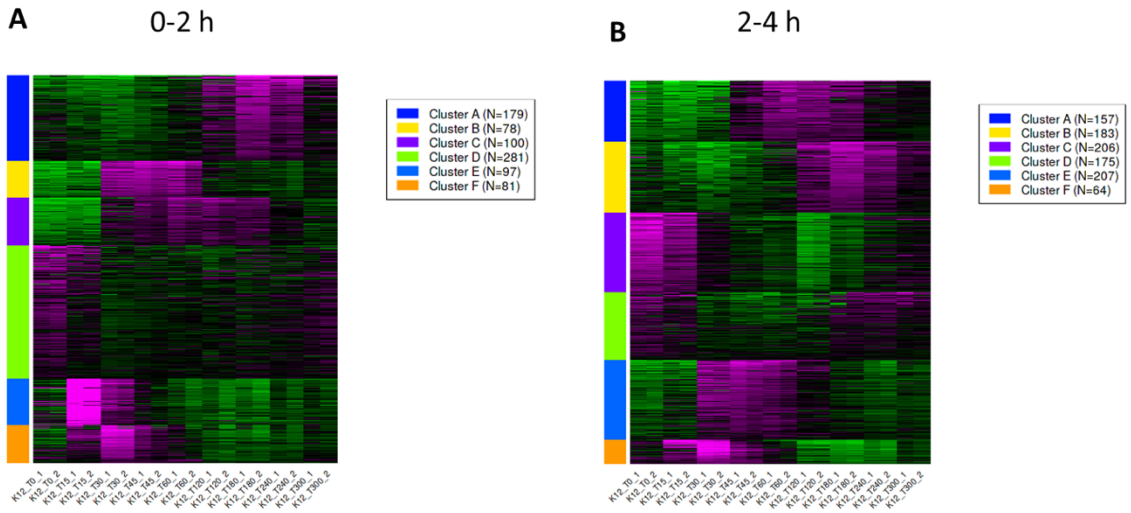

Figure 23- *k*-means clustering to 6 clusters of differentially expressed genes from Experiment 3 mapped onto Experiment 1 **A-** genes from Experiment 3 that were differentially expressed from 0-2 hour timepoints, **B-** genes from Experiment 3 that were differentially expressed from 2-4 hour timepoints.

To further investigate the lack of synchrony between Experiments 1 and 3, as was done for Experiment 2 (Figure 19), differentially expressed genes from Experiment 3 were taken and their expression patterns interrogated in Experiment 1. Figure 23 shows *k*-means clustering of gene expression in Experiment 1, for the subset of genes that are significantly differentially regulated between 0-2 hours (A) and 2-4 hours (B) in Experiment 3. This confirms, that as for Experiment 2 (Figure 19), these genesets from specific timepoints in Experiment 3 are expressed throughout the predation cycle in Experiment 1, confirming that predation in Experiment 3 is asynchronous relative to Experiment 1. However, there was some significant co-expression of genes at the timepoints between Experiments 2 and 3, as seen above, and, for example there were 449 genes called as significantly differentially regulated from time 0 to 2 common to Experiments 1 and 2. Therefore the comparisons discussed above are valid and form a solid basis for future investigations of the genes, although it should be borne in mind

that some differences may have also arisen from the lack of synchrony between Experiments 2 and 3 relative to Experiment 1.

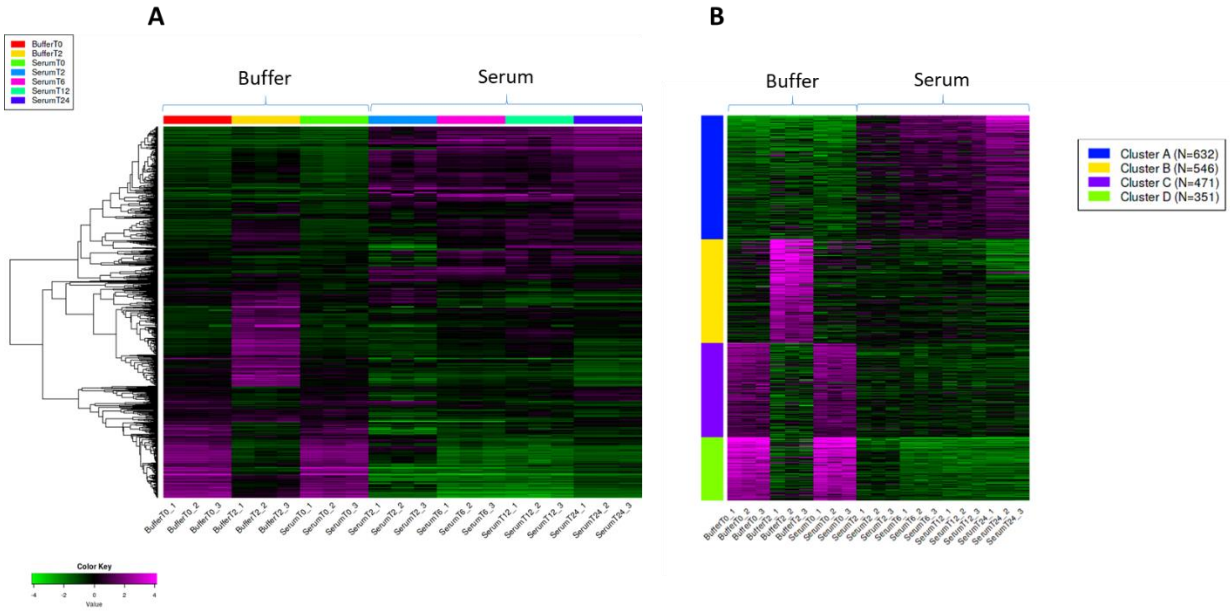

Figure 24 A- dendrogram of hierarchical clustering for samples of *B. bacteriovorus* alone in buffer or in serum **B-** *k*-means clustering to 4 clusters for these.

In order to determine the response of the *B. bacteriovorus* cells to serum, these were incubated in the absence of prey in buffer control and in serum, for up to 24 hours and their transcriptional response was measured. Figure 24 shows clustering analyses of these samples.

Clusters C and D represent the genes expressed in attack phase at the start of the experiment and consist of general metabolic housekeeping genes, with Cluster D enriched for gene expression and translation and protein metabolism (as identified above in Cluster G in Figure 20B). These are virtually identical in buffer and serum as samples were collected immediately after resuspending in the medium, with no time for expression to be altered.

Cluster B consists of genes upregulated upon prolonged incubation in buffer, but not in serum. This cluster contains a lot of non-coding RNAs (likely antisense, suppressing translation of genes in this nutrient-free environment) and lots of genes of unknown function. There are also several flagellar motility and pilus related genes.

Cluster A is the most interesting as this consists of genes that are upregulated progressively more as incubation in serum is prolonged, with genes in this cluster most highly expressed after 24 hours in serum. These include siderophores and other iron uptake mechanisms, outer membrane proteins, various transmembrane transport systems and some pilus-related genes. Although this cluster also includes various proteases, it is interesting to note that this very rich medium does not seem to upregulate growth or replication mechanisms. Instead, the medium contains potentially hostile antibacterial factors and this is reflected in what is potentially a defence reaction by *B. bacteriovorus,* with some DNA defence mechanisms also in this cluster.

Conclusions from Experiment 3

As with Experiment 2, where predation on *Serratia* did not appear to be occurring synchronously as it was for predation on *E. coli*, this was true for predation in serum as it was in buffer. Many genes were regulated specifically for predation on the different prey, but also in the different media, with the *B. bacteriovorus* expressing a range of genes in response to the challenging medium of serum. This was particularly noteworthy at the 4 hour timepoint, where the serum changed as a medium and the *B. bacteriovorus* responded to this with a global transcriptional response. The lack of synchrony is likely due to the switching of prey from *E. coli* to *Serratia*; we have previously observed that predation is not immediately as efficient when changing prey. The mechanism for this is currently unknown, but when *B. bacteriovorus* is grown continuously on the same prey, then predation on that same prey again is more efficient. However, if *B. bacteriovorus* were to be used as a therapeutic, it would be necessary to grow it first on a non-pathogenic prey before applying to a clinical infection, hence the design of the experiment here with growth first on harmless *E. coli*. The results here show that the effect of this is that the predation on the clinically relevant strain in non-synchronous, presumably with some *B. bacteriovorus* entering prey early and some later, so that predation of the population was at all different stages, as seen by mapping the genes differentially regulated on *Serratia* onto Experiment 1.

The reaction of *B. bacteriovorus* to serum was apparently one of defense against the potentially harmful substances within this medium. Instead of inducing degradative enzymes and general growth and metabolism genes as happens with growth in nutrients (see Experiment 6 below), various outer membrane proteins are produced along with transport systems which are potentially exporting toxins (e.g. paraquat induced transmembrane transporter is upregulated which in other bacteria exports the toxin from its cell). Further, DNA protecting systems are also upregulated. Whilst there are some proteases expressed, these do not appear to be the main degradative proteases identified (see Experiment 5 below). This reaction seems to happen in serum with and without the prey as the same systems were identified in both conditions. The fact that *B. bacteriovorus* does not seem to metabolise and grow in human serum bodes very well for its use as a therapeutic as either of these traits could cause problems in a clinical setting.

Experiment 4- predation by *B. bacteriovorus* HD100 mutant *ΔdgcC* on *Serratia marcescens* in buffer and pooled human serum samples throughout the predation cycle.

Experiment 4 is essentially identical to Experiment 3 except predation throughout the timepoints is with the mutant *ΔdgcC* strain in both buffer and serum. The mutant strain has a deleted global regulator gene *ΔdgcC*, the product of which is a GGDEF cyclic-di-GMP producing enzyme which controls the ability to switch to Host-Independent (HI) growth. The hypothesis is that the strain's inability to switch on HI growth may result in a dedicated predator, which would prey effectively and not be able to persist without prey; a desirable outcome for use as a clinical therapy. Alternatively, the lack of an important global regulator may hinder normal functional regulation and therefore predation.

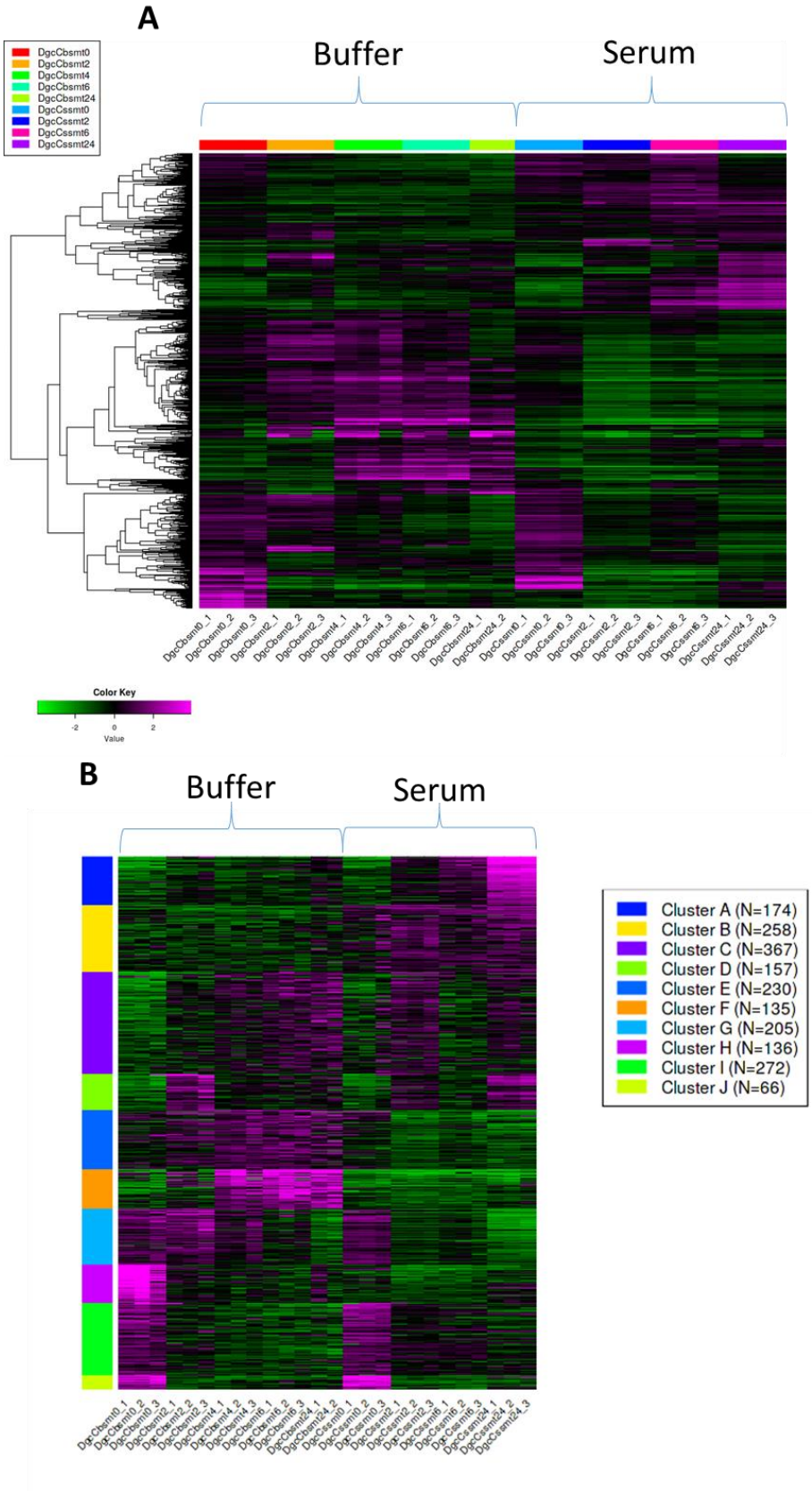

Figure 25 A- dendrogram of hierarchical clustering for Experiment 4 **B-** *k*-means clustering to 10 clusters for Experiment 4.

Figure 25 shows that results for Experiment 4 are similar to those of Experiment 3 (Figure 20), with predation of the mutant strain *ΔdgcC* appearing not to be synchronous on *Serratia*, in neither buffer nor serum. Similar pathways were enriched in the same way as for the predation with the wild-type predator, for example energy and nucleotide metabolism enriched in cluster C, expressed at a variety of timepoints throughout the predation cycle. Unfortunately, there was not a 4 hour timepoint for this experiment in serum to see if the same response to changes in serum was also present in this mutant. However, as with the wild-type strain, there were clearly many genes that were specific to predation/survival in serum in Clusters A and B.

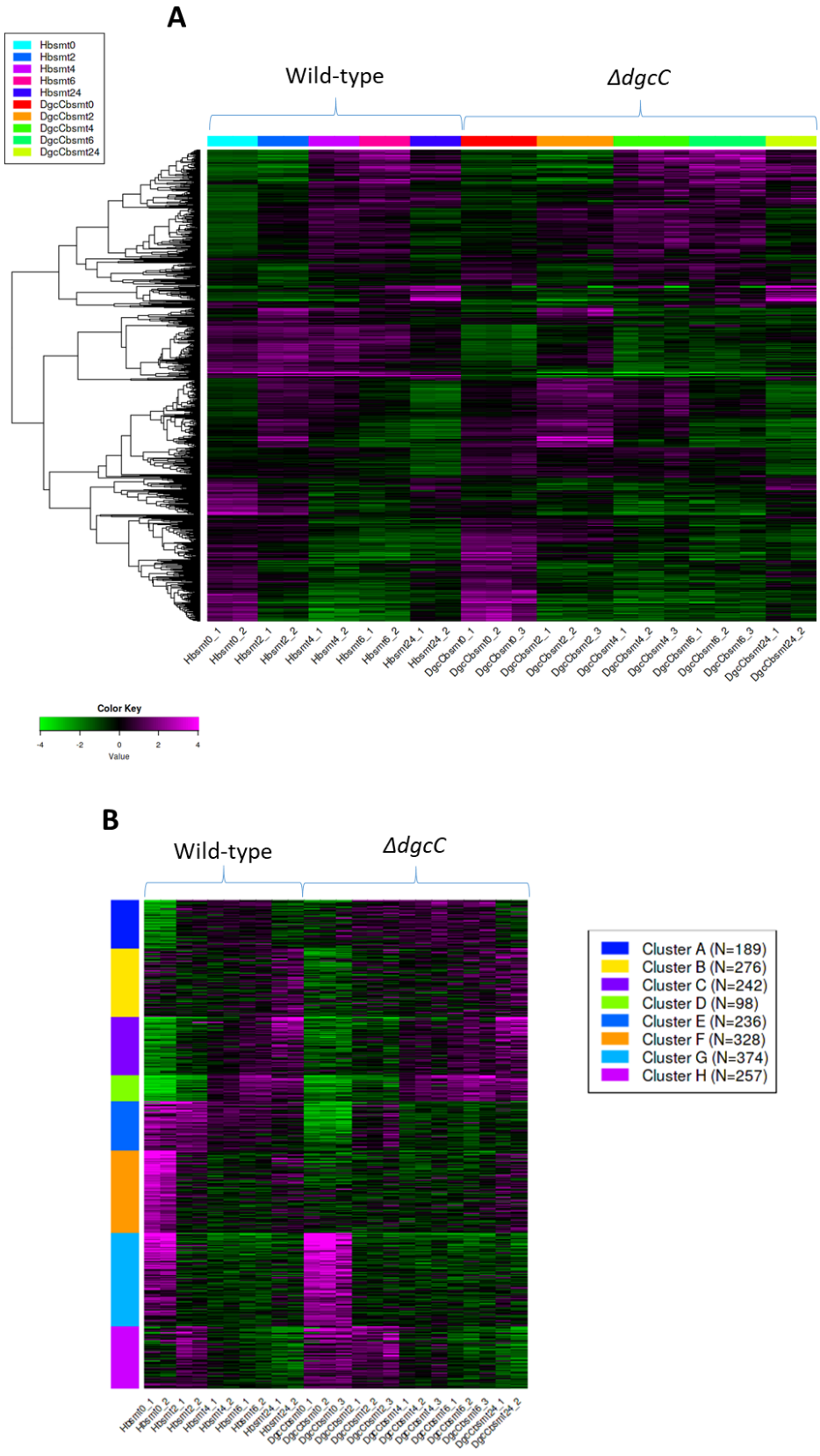

Figure 26 A- dendrogram of hierarchical clustering for experiments in buffer **B-** *k*-means clustering to 8 clusters for experiments in buffer.

Figure 26 compares expression of wild-type versus Δ*dgcC* mutant throughout the predation cycle in buffer. On the whole, most genes are showing similar expression patterns in the wild-type and the mutant with pathways for energy and nucleotide metabolism enriched in Cluster B, peaking expression at 24 hours. Interestingly, DNA replication pathways are enriched in Cluster C, which is also expressed very late on, peaking at 24 hours. Both of these suggest that the predation on *Serratia* may be delayed compared to *E. coli* as these processes peak at 60-120 mins with predation on *E. coli* (Figure 14).

The biggest clear difference between the strains is the massive global upregulation of genes at time 0 in the wild-type that is not present in the mutant in Cluster F. This Cluster is enriched for pathways in gene expression and translation, general metabolism, response to nitrogen compounds and RNA processing. This is likely housekeeping attack phase genes turned on in the wild-type but not in the mutant.

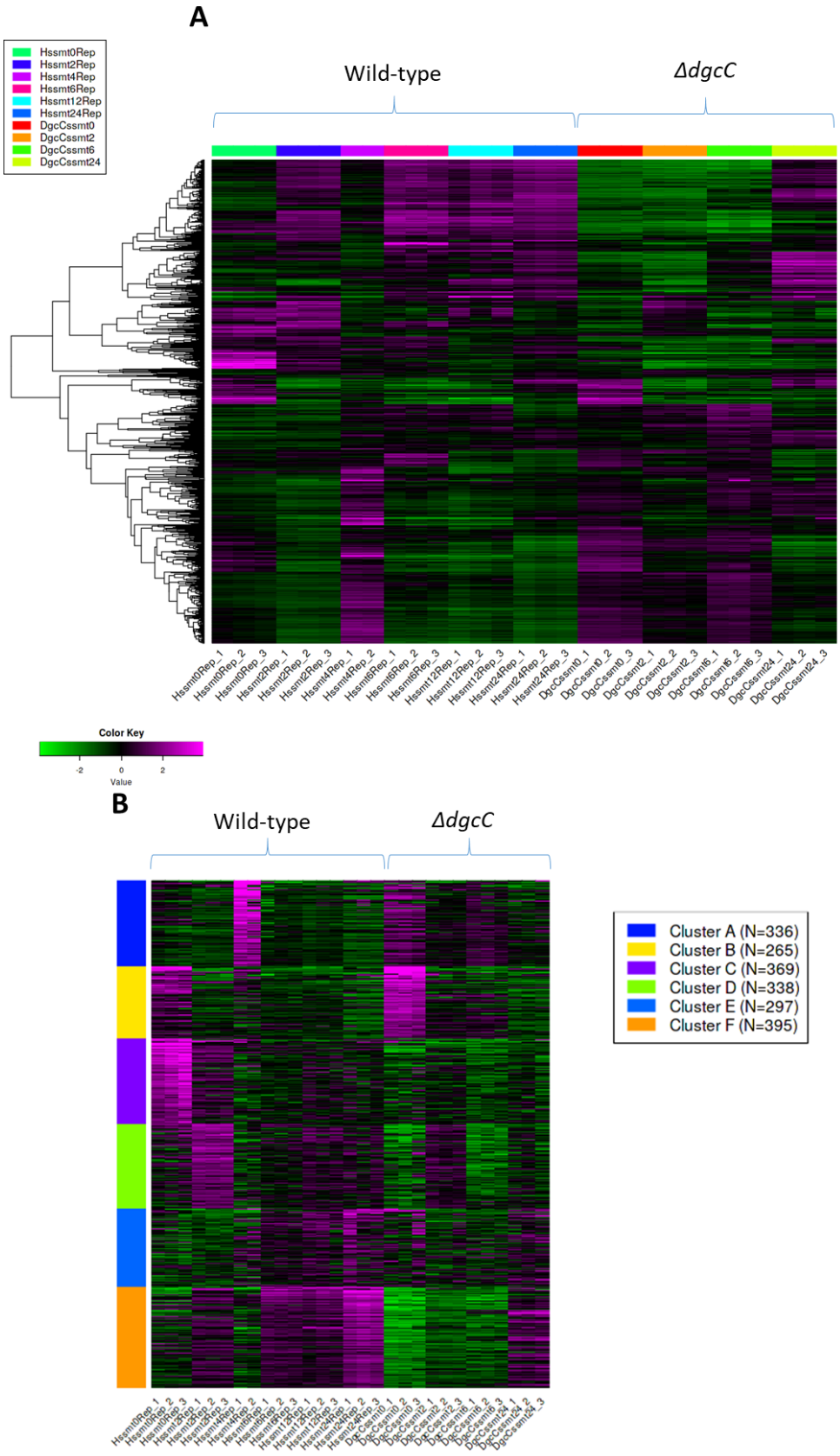

Figure 27 A- dendrogram of hierarchical clustering for experiments in serum **B-** *k*-means clustering to 6 clusters for experiments in serum.

Figure 27 compares expression of wild-type versus Δ*dgcC* mutant throughout the predation cycle in serum. There are again many similarities between expression patterns in the mutant and wild-type, but several differences. The first obvious difference is that of Cluster A, with lots of genes highly upregulated at 4 hours in the wild-type. Unfortunately, there is no sample for the mutant at this timepoint, so this may also be true for the mutant, however we can see that many of these genes are expressed at both 0 and 6 hours in the mutant and that this is not true for the wild-type. Again, the biggest apparent difference between the strains was Cluster C with pathways for energy and nucleotide metabolism enriched (equivalent of Cluster B in Figure 26), which are highly upregulated early in wild-type, but not in the mutant. In serum, these continue to be expressed at 2 hours, but less so in buffer reaffirming the observed difference between the media. Genes in Cluster D were enriched for pathways in energy and phosphate metabolism as well as ion transport and whilst these were upregulated at 2 hours with both predators, many were still expressed at later timepoints in the wild type, but less so in the mutant. Similarly, Cluster E were expressed at 24 hours in both wild-type and mutant, but were also expressed at 6-12 hours in the wild-type and expressed at 0 hours in the mutant.

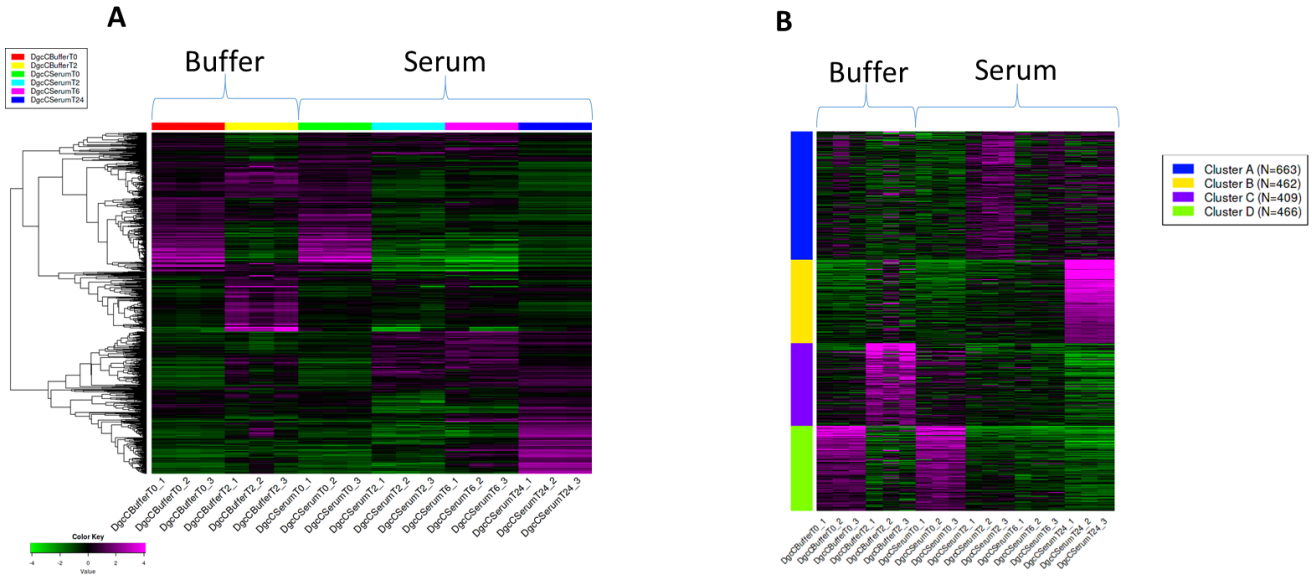

Figure 28- dendrogram of hierarchical clustering for samples of *B. bacteriovorus ΔdgcC* mutant alone in buffer or in serum **B-** *k*-means clustering to 4 clusters for these.

As for the wild-type, in order to determine the response of the *ΔdgcC* mutant cells to serum, these were incubated in the absence of prey in buffer control and in serum, for up to 24 hours and their transcriptional response was measured. Figure 28 shows clustering analyses of these samples.

As for the wild-type, Cluster D represents attack phase genes and includes general metabolism genes, but is also enriched for genes involved in nitrogen compound metabolism. Again, as for the wild-type there is virtually no difference between buffer and serum at the first timepoint as both were immediately harvested.

Cluster C represents a response to incubation in a nutrient-free environment and was similar to the response in the wild-type with lots of non-coding RNAs and flagellar motility and pilus related genes.

However, the response of the Δ*dgcC* mutant to serum was markedly different to that of the wild-type. Cluster A consists of genes upregulated after 2 hours of incubation in

serum, but downregulated upon further incubation. This includes a large number of genes of unknown function, but also includes some genes upregulated later in the wild-type, such as paraquat induced transporter and some DNA protection genes. Also in the cluster are the alternate sigma factor *rpoE*, chaperone *groES* and global regulator *fis*, suggesting that there is a large scale dyregulation of genes in the Δ*dgcC* mutant.

Cluster B contains the Δ*dgcC* mutant's main response to serum, which unlike the wild-type, occurs only after 24 hours in serum. This includes a great many flagellar motility genes and pilus generating genes. In order to compare the wild-type and mutant response, the data were merged and analysed.

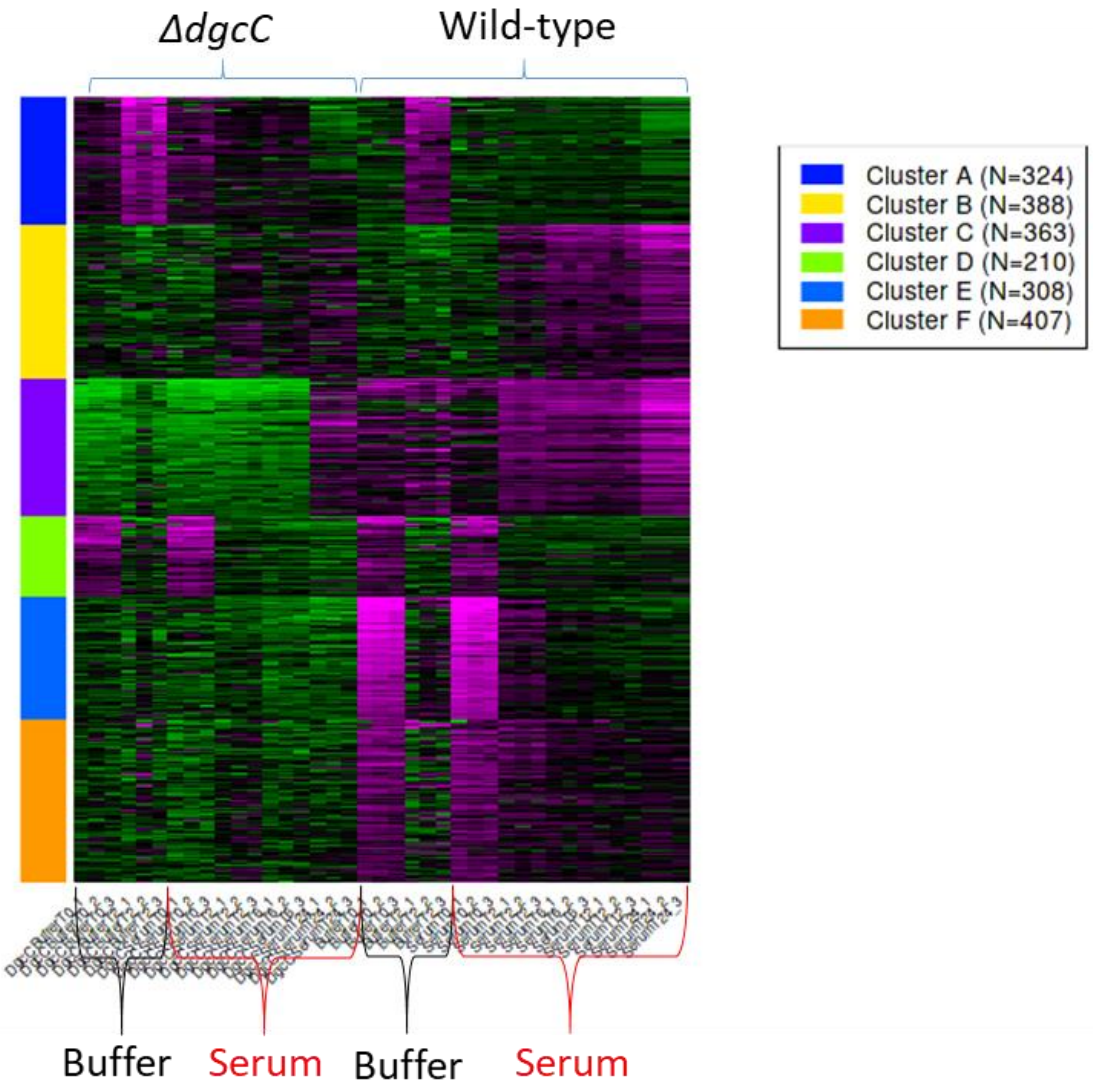

Figure 29- *k*-means clustering to 6 clusters for samples of *B. bacteriovorus ΔdgcC* mutant or wild-type alone in buffer or in serum.

Figure 29 compares the response of the Δ*dgcC* mutant to that of the wild-type to incubation in buffer or serum. Cluster A consists of genes upregulated upon incubation in buffer for 2 hours. The response is similar for the wild-type and the Δ*dgcC* mutant, with lots of non-coding RNAs, several flagellar motility and pilus forming genes and the alternative sigma factor E encoded by *bd0881*, known to be involved in flagellar motility regulation (Lambert et al., 2012).

Cluster B is also a response that the mutant and wild-type have in common, but to prolonged incubation in serum. The response is stronger (in that the genes are more upregulated) in the wild-type compared to the mutant. The cluster includes genes for pilus assembly, iron scavenging, transmembrane transporters, DNA protection and competence.

Cluster C contains genes that are upregulated in response to serum. In the wild-type many of these are also expressed in buffer and the response to serum is fairly strong with just 2 hours incubation in serum, growing stronger with longer exposure. In the Δ*dgcC* mutant, however, there is virtually no expression of these genes except for in the sample incubated with serum for 24 hours. The main feature of this cluster is a great number of genes involved in flagellar motility and several involved in pilus formation. Interestingly, the cluster also contains some genes annotated as involved in gliding motility, but no full operon for this is included. Similarly, some *fts* genes are in this cluster, but again, not a full complement to enable cell division.

Cluster D consists of genes highly expressed in both mutant and wild-type in attack phase at time 0, in both buffer and serum. These include alternate cytochromes, catalase genes involved in oxidative shock response and also includes the operon *bd2224-2229* thought to be involved in early predation signalling.

Both Clusters E and F are genes that are expressed in attack phase in the wild-type, but not in the mutant Δ*dgcC*. Cluster E is enriched in genes for transcription and translation, protein and nitrogen compound metabolism, and macromolecule synthesis suggesting that these are general housekeeping genes. Here, a different alternative sigma factor E, *bd0743*, and alternative chaperone *groES* are expressed.

<u>Conclusions from Experiment 4</u>

Experiment 4 largely confirmed the results from Experiment 3 by demonstrating that predation on *Serratia* did not result in clear temporal expression of known functional genes throughout the predation cycle as was the case for predation on *E. coli*. Similarly, it confirmed that there is a significant subset of genes upregulated only in the presence of serum and not buffer. This set of genes could be vital for future research in the use of *B. bacteriovorus* as these could determine if the predators are capable of surviving in clinically relevant settings and for how long. This balance is important to understand as ideally, the predators would need to evade bactericidal elements (such as complement in serum) for long enough to enter target prey and be effective at killing them, but not be too resistant that the predators would themselves persist to potentially become a hazard.

There were also clear differences in gene expression between the wild-type and the mutant, with many groups of genes being expressed at different timepoints. Most notable was a global induction of genes at a very early point in the wild-type that was absent in the mutant. DgcC is a global regulator which, when absent in the Δ*dgcC* mutant prevents this from being able to switch to host-independent (HI) growth (Hobley et al., 2012a). The switch to HI growth occurs at low frequency and is usually a result of a mutation in the *bd0108* gene, although not exclusively (Capeness et al., 2013). In nature, signals (as yet unknown) may result in inducing HI growth without mutation. HI growth then utilises nutrients from the environment for growth and division, expressing many of the genes involved in these processes throughout the growth phase of predation (30-180 mins for the *E. coli* Experiment 1). It may be that such a signal is present to some extent in attack phase, but that the Δ*dgcC* mutant is defective in reacting to this, which is why the initial expression of these genes seen in the wild-type is absent in this

mutant. This may then lead to the mutant and wild-type predation becoming out of synch with each other and explain the different temporal expression of so many genes.

Analysis of the *B. bacteriovorus* alone in both buffer and serum revealed significant differences between the mutant and wild-type strain suggesting that the mutant has a drastically changed global gene expression pattern, with large amounts of housekeeping genes turned off in attack phase and flagellar gene expression upregulated at very different times to the wild-type. This analysis further suggested that the differences in gene expression in serum were similar in the mutant and wild-type, but with delayed response in the mutant. The response to serum seemed essentially the same in the presence or absence of prey, and seemed to be a defensive response rather than a growth response to the rich nutrients. Genes for iron scavenging, transmembrane import/export and DNA protection were induced rather than digestive enzymes and metabolism and growth pathways, which are induced in the presence of nutrients (see Experiment 6 below). This suggests that *B. bacteriovorus* first and foremost responds to the antibacterial elements of the serum. That serum does not induce growth and replication in *B. bacteriovorus* is a very important discovery that bodes well for its potential as a therapeutic as unrestrained growth of the predator may cause problems.

Experiment 5- Interaction of *B. bacteriovorus* HD100 with the Gram-positive *Staphylococcus aureus*.

B. bacteriovorus is an intracellular predator of other Gram-negative bacteria, growing within the prey periplasm and cannot prey upon Gram-positive bacteria which lack periplasms. However Im and co-workers (Im et al., 2018) noted that Gram-positive *Staphylococcus aureus* biofilms were dispersed to some degree by the presence of *B. bacteriovorus* with a concomitant reduction in *Staphylococcus* numbers and increased viability of *B. bacteriovorus*. To study this interaction, they isolated RNA from *B. bacteriovorus* after incubation with biofilms of *S. aureus* and *B. bacteriovorus* controls in buffer alone.

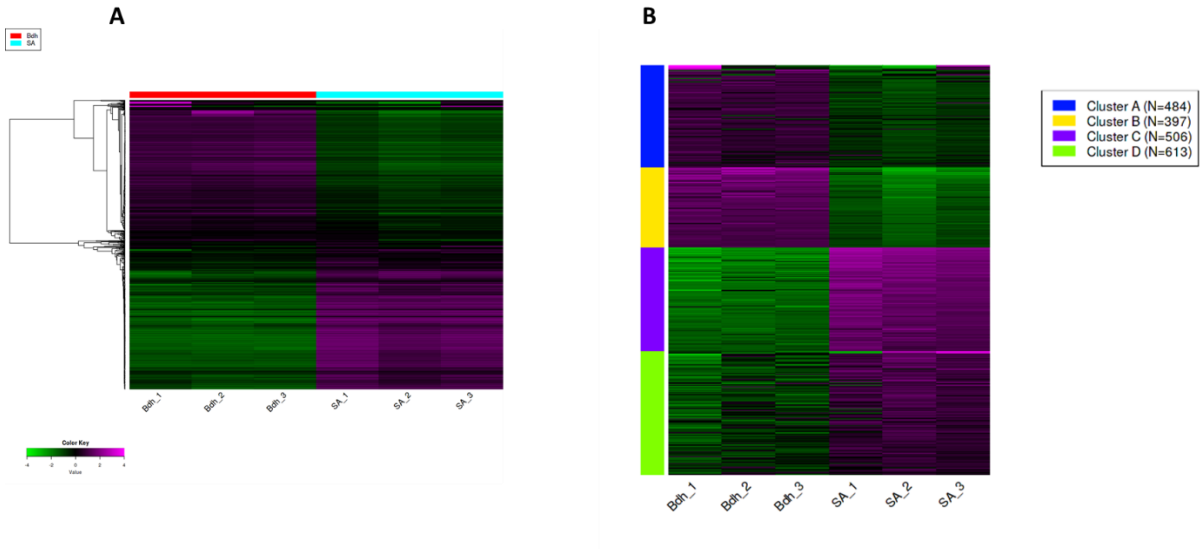

Figure 30 A- dendrogram of hierarchical clustering for Experiment 5 **B-** *k*-means clustering to 4 clusters for Experiment 5. Bdh, *B. bacteriovorus* alone in buffer control, SA, *B. bacteriovorus* with *Staphylococcus aureus* biofilms.

Figure 30 shows the gene expression of *B. bacteriovorus* in response to *S. aureus* biofilm, compared to a buffer control. As expected for 2 conditions, all differentially regulated genes are either up or down regulated in each condition. Clustering to 4 clusters (Figure 30B) shows there are groups with significantly different levels of expression. Cluster D is genes upregulated upon contact with biofilm and is enriched for pathways in gene expression and translation as well as protein and nitrogen metabolism. The authors concluded therefore that gene expression in response to biofilms was analogous to response to nutrients and intraperiplasmic growth. To investigate this further, expression in response to biofilms was compared to the whole cycle of intraperiplasmic growth.

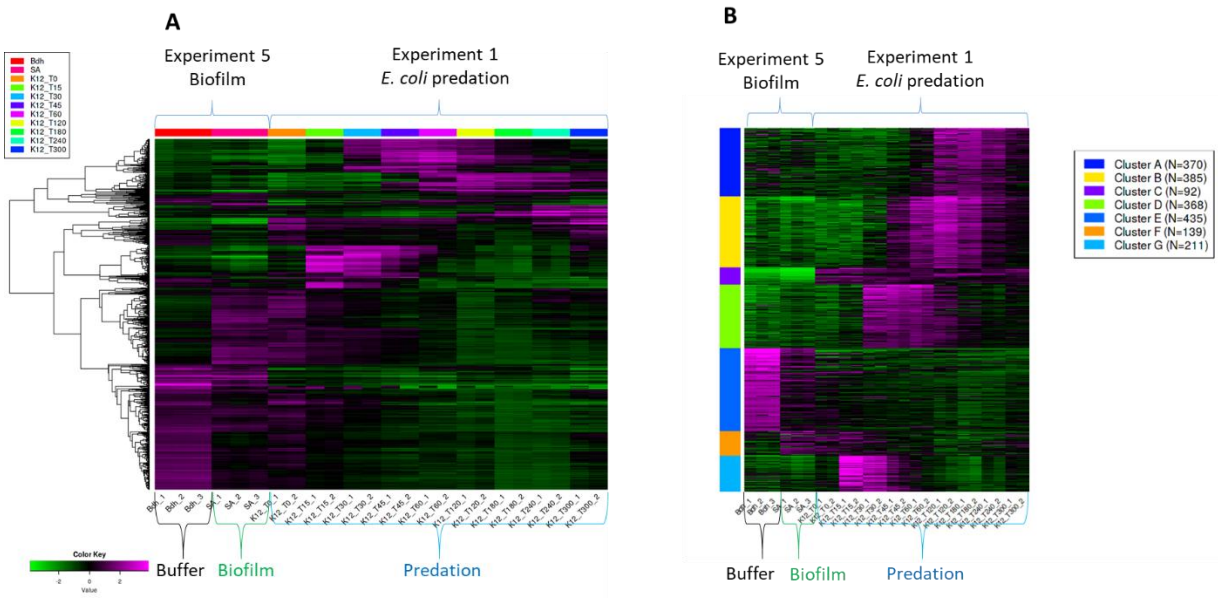

Figure 31 A- dendrogram of hierarchical clustering for Experiments 1 and 5 **B-** *k*-means clustering to 7 clusters for Experiments 1 and 5.

Figure 31 shows a comparison of expression throughout the predation cycle with the response to *S. aureus* biofilms and this suggests that combining these experiments is invalid as there are apparently virtually no genes upregulated upon contact with biofilm, which Figure 30 (and the authors' analysis) shows to be wrong. The reason for this is likely the different conditions in which the experiments were undertaken (with slightly different growth concentrations and conditions, and different sequencing depths) and the widely varying levels of expression in the different conditions. Therefore, this attempt to normalise the expression across the two experiments resulted in anomalies. For example, the genes *bd2269* and *bd2692* were examined as significantly upregulated upon biofilm exposure and their products tested to be involved in biofilm dispersal, but because of much higher levels of expression through the predation cycle, the difference in the biofilm versus its control was obscured. The same occurred with most of the biofilm-associated genes and therefore this comparison is not useful. This reiterates the importance of adequate controls in experiments and the importance of consistency in experimental conditions for techniques as sensitive as RNA-Seq. Because of this, an alternative approach was taken to analyse this dataset: the upregulated gene set from interaction with biofilms (Clusters C and D from Figure 30B) were taken and examined for their expression pattern in Experiment 1.

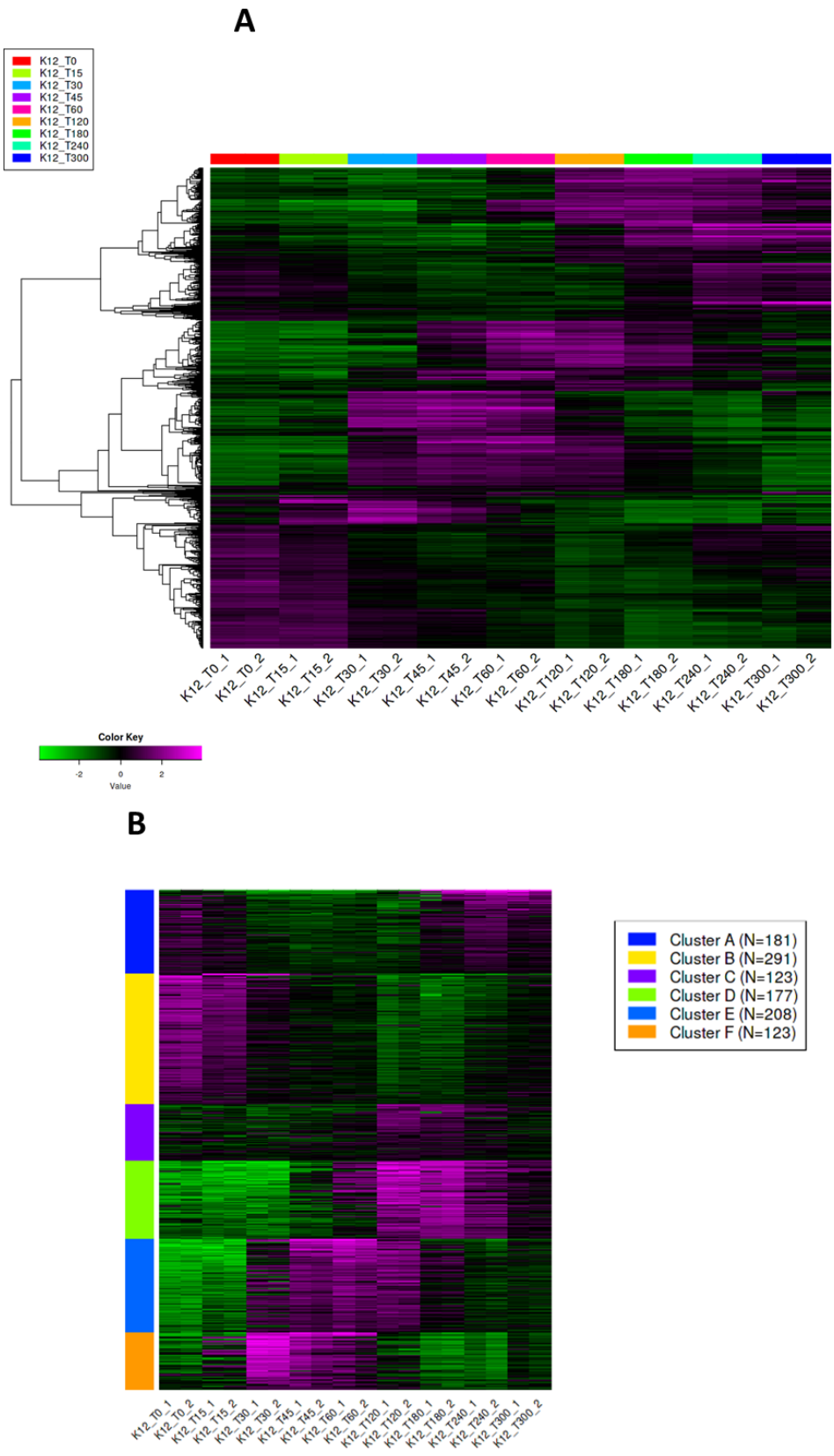

Figure 32 A- dendrogram of hierarchical clustering for genes from Experiment 1 that were upregulated in contact with *S. aureus* biofilms in Experiment 5 **B-** *k*-means clustering to 6 clusters for these genes.

Figure 32 shows hierarchical clustering and *k*-means clustering from Experiment 1 for the geneset that was upregulated in Experiment 5 upon contact with *S. aureus* biofilms. This shows that this alternative analysis approach is valid, giving clusters of genes that are upregulated at different times throughout the predation cycle in Experiment 1, in agreement with the broad conclusions drawn by the authors of this work (Im et al., 2018).

Cluster A are genes highly expressed at time 0 and later in the lifecycle as new attack phase cells are being released (T240-300) and these consist of genes known to be associated with attack phase cells, including motility and pilus related genes and gene whose proteins are in large abundance such as outer membrane proteins (OMPs) and Tol proteins. These were to be expected as the *B. bacteriovorus* in the experiment in response to *S. aureus* biofilms appeared to still be in attack phase rather than growing and dividing.

Cluster B are genes highly expressed at time 0, but that also stay on during the earlier interactions with prey cells. These include lots of flagellar motility, chemotaxis and pili-related genes.

Interestingly, both Clusters F and C are highly enriched for many RNA and DNA metabolic processes, but these are expressed exclusively near the start (Cluster F) or near the end (Cluster C) of predation. Because of these expression patterns, it is believed that those in Cluster F are predominantly involved in prey nucleotide degradation and that those of Cluster C are predominantly *B. bacteriovorus* nucleotide metabolism. It's interesting to note that these two distal groups are co-expressed in contact with *S. aureus* biofilms. Also in Cluster C is *bd0125*, which is annotated as *comL*, a competence-related gene for DNA uptake.

Cluster E consists of genes upregulated during *B. bacteriovorus* intraperiplasmic growth (30-120 minutes) and is enriched with pathways for gene expression and translation and protein and nitrogen metabolism. This also includes proteases and siderophores *bd1574-1576, iucCBA* and *bd0070 hemK* for iron uptake.

Cluster D consists of genes upregulated later during intraperiplasmic growth and division, but does not seem to contain many genes associated with genome replication and cell division (although *ftsZ* is in this group, very few genes required for DNA replication and cell division are present agreeing with the observation that these processes are not happening). Rather, the Cluster consists of genes associated with protection, such as DNA binding *bd0025-6 sdhAB*, *bd0254 uvrC* and *bd1954 mutS.* In this Cluster, there are also purine metabolism genes and proteases, including *bd2269*, identified and further studied by the authors.

<u>Conclusions from Experiment 5</u>

One of the conclusions from the analysis of Experiment 5 is that it is important to test methods of comparisons as normalising the data in combination with data from Experiment 1 did not give valid results. The bacteria were grown with slightly different methods, although in both experiments they were grown in Ca/HEPES buffer on *E. coli* prey, the concentration and condition of the prey cells upon which the predatory cultures were originally grown differed (fresh overnight grown in Experiment 1, concentrated, washed and stored at 4°C for Experiment 5). It is likely that even this minor detail resulted in significant differences as RNA-Seq is a very sensitive technique. This could be an important observation in trying to get standardised methods for predator growth across different research groups.

However, analysing the expression in Experiment 1 of the genes upregulated in contact with *S. aureus* biofilms revealed in greater detail their potential role. The authors noted that the *B. bacteriovorus* cells interacting with the biofilm were attack phase cells, did not differentiate into growing, longer cells (for example as happens to host-independent mutants) and did not enter the Gram positive *S. aureus* cells, thus must be eliciting a very different response to contact with prey cells. Many genes upregulated in attack phase were upregulated in contact with biofilms as expected by the observation that the cells appeared to be attack phase (short, and highly motile). However, it is interesting to note that the cluster of genes that stay upregulated upon early contact with prey (Cluster B) are also remain upregulated upon contact with the very different Gram positive *S. aureus*. This is in agreement with microscopic observations that initial contact by *B. bacteriovorus* is non-specific, with contact with inanimate surfaces, and indeed non-prey such as *S. aureus* happens briefly, before contact is abandoned and swimming resumes. This further suggests a role in some of the early expressed genes in contact with valid Gram-negative prey are important in secondary, prey-specific contact important for prey entry and predation. The genes in these clusters (A and B) further show the importance of flagellar motility and chemotaxis for the attack phase *B. bacteriovorus* and suggest that this is also vital for their interaction with biofilms.

The other clusters agree with the authors' conclusions that general nutrient usage and metabolism is induced in the absence of predator growth and division. These also show that DNA and RNA uptake and metabolism is occurring, that iron scavenging and protein breakdown and uptake are happening. The upregulation of DNA protection agents may reflect a defence mechanism against rapid nucleotide uptake from harming the predator's own DNA, something that also may be happening during intraperiplasmic growth.

Experiment 6- *B. bacteriovorus* response to nutrients

B. bacteriovorus attack phase cells are incapable of DNA replication and cell division without acquiring a mutation (usually, but not exclusively, at the *bd0108* locus (Capeness et al., 2013)) to turn them in to host independent (HI) mutants. As such, they are incapable of growth in normal lab media, but several reports suggest that they benefit from external nutrients without growing. Experiment 6 set out to address this by comparing attack phase cells in nutrient-free Ca/HEPES buffer with nutrient rich broth (1 x NB).

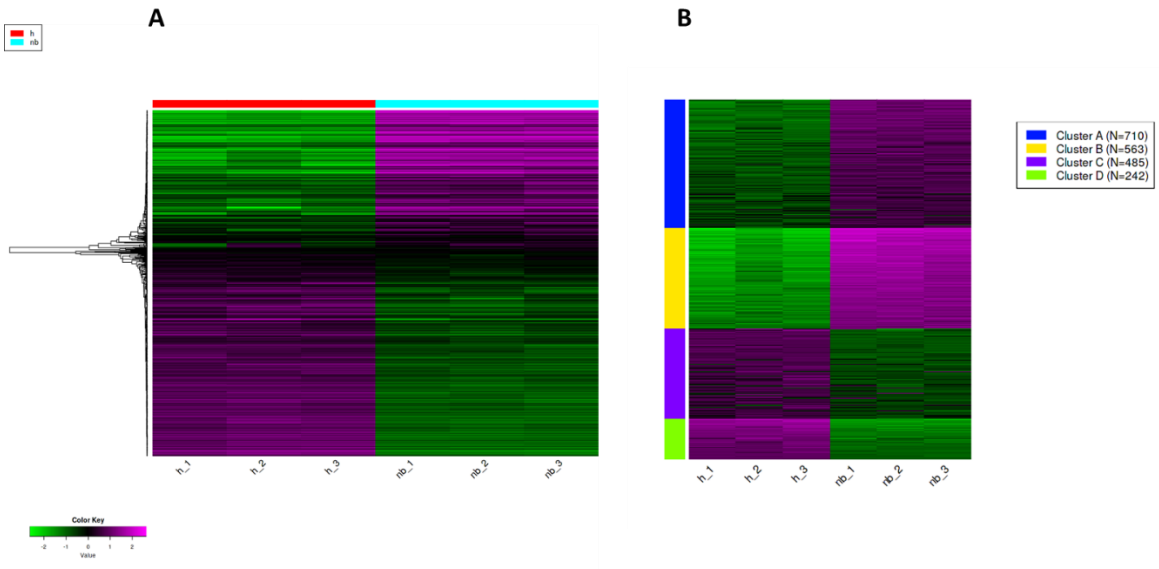

Figure 33 A- dendrogram of hierarchical clustering for Experiment 6 **B-** *k*-means clustering to 4 clusters for Experiment 6. h, *B. bacteriovorus* alone in buffer control, nb, *B. bacteriovorus* incubated with nutrient broth.

Figure 33A shows as expected that for an experiment with 2 conditions, hierarchical clustering results in a group of genes upregulated and a group of genes downregulated, whilst Figure 23B shows that these can be clustered into groups with significantly different expression levels.

Experiment 6 was carried out by the same research group as Experiment 5, using the same growth conditions for the *B. bacteriovorus* and similar experimental conditions (such as cell concentrations). Therefore attempts to normalise the dataset along with the data from Experiment 1 also did not give valid results for this experiment. Therefore, again the alternative approach was used of extracting the upregulated genes in response to nutrients and testing their expression patterns in Experiment 1 to determine their functions.

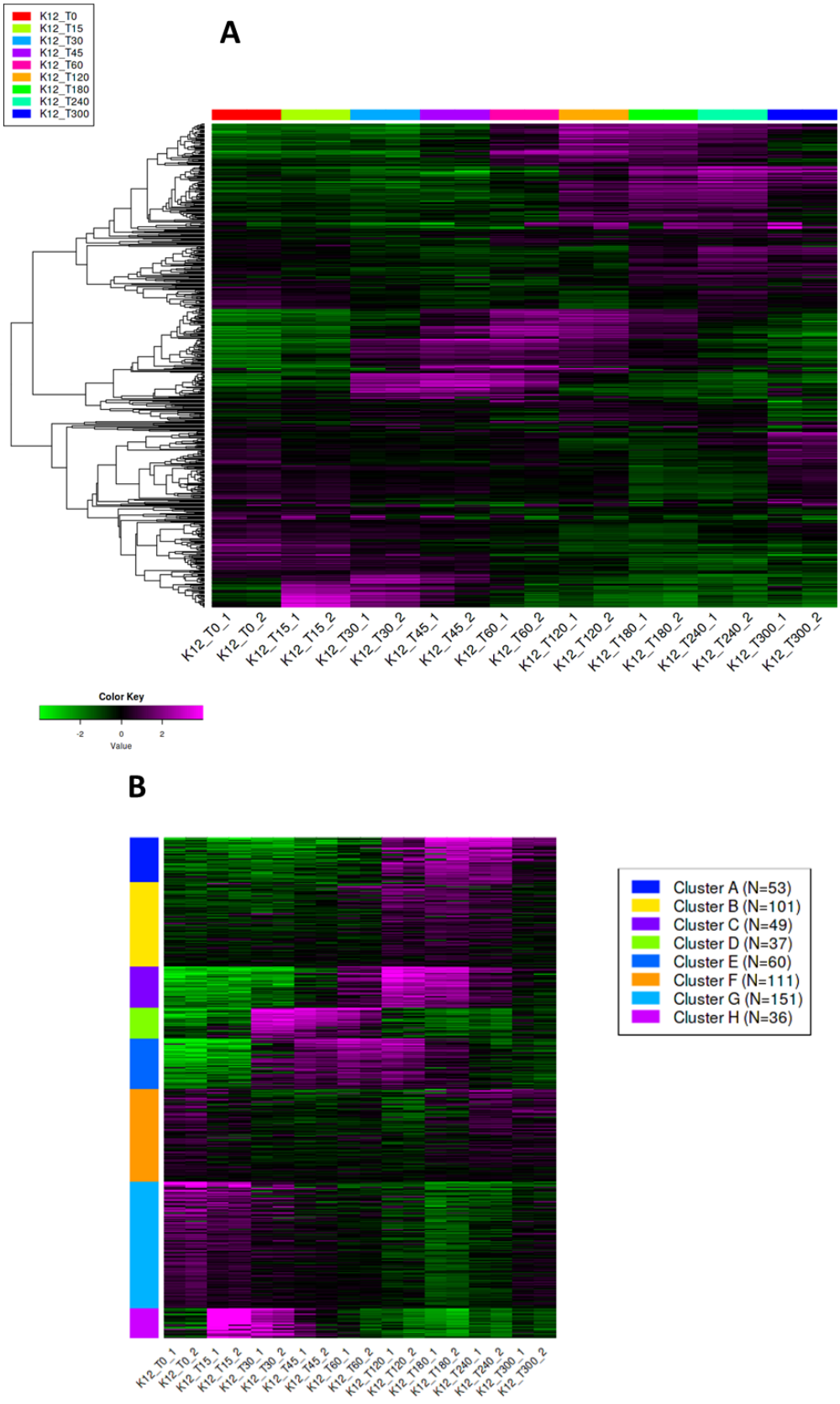

Figure 34 A- dendrogram of hierarchical clustering for genes from Experiment 1 that were upregulated in response to nutrients in Experiment 6 **B-** *k*-means clustering to 8 clusters for these genes.

Figure 34 shows that genes upregulated in response to nutrients in Experiment 6 are expressed at all different times in the predation lifecycle in Experiment 1. Cluster E is the only cluster which is enriched for annotated genes and this represents genes expressed throughout the growth phase of the predatory lifecycle (from 45-180 mins). These include genes for transcription, translation and protein and nitrogen metabolism.

Clusters A and B are genes expressed at the end of the predatory cycle, when new attack phase cells are generated and released from the bdelloplast. Interestingly, within this group are highly abundant proteins such as OMPs with would need to be synthesised in large amounts at this stage in the predatory cycle. The authors noted that the cells exposed to nutrients were significantly longer than the control cells so it seems they are using these genes for cell growth. However, the authors did not see any cell division or drastic increase in viability associated with cell proliferation. Intriguingly, both the divisome-associated genes *ftsA* and *ftsZ* and the gene associated with cell wall generation; *pbpC* are within Clusters A and B, but there is not a complete set of divisome genes so it may be that these gene products have a role in regulating the growth of the cell wall without necessarily initiating cell division.

Cluster C is genes highly upregulated just before this stage (60-240 mins, peaking at 120) and this includes the cytoskeleton genes *mreB* and *mreC*, which again, may be in preparation for cell growth.

Cluster D consists of genes upregulated earlier (30-60 minutes) and includes the extracellular nuclease-encoding *bd1934.* Nutrient broth does not contain significant amounts of polynucleotides, so it seems likely that arsenals of hydrolytic enzymes are under global transcriptional control, with some being turned on without necessarily being in conditions for their use.

Clusters F represents attack phase genes only, highly expressed at time 0 and 300 mins and Cluster G represents attack phase genes that remain on upon early contact with prey. These clusters again contain genes associated with flagellar motility, chemotaxis, pili and outer membrane components such as *tol.*

Whilst small, Cluster H is highly unexpected as these are genes highly upregulated upon prey attachment. Most are unannotated, but from our work, we can see several genes that encode L,D-transpeptidases: *bd0553, bd1176,* and *bd3176*. As these are thought to be exported into the bdelloplast to act upon the prey cell wall (and for Bd1176, this has been proven (Kuru et al., 2017)), it seems that something in the nutrient broth is acting as a secondary signal normally present in prey to induce expression of these genes, whose products are very unlikely to be useful in utilising the nutrients.

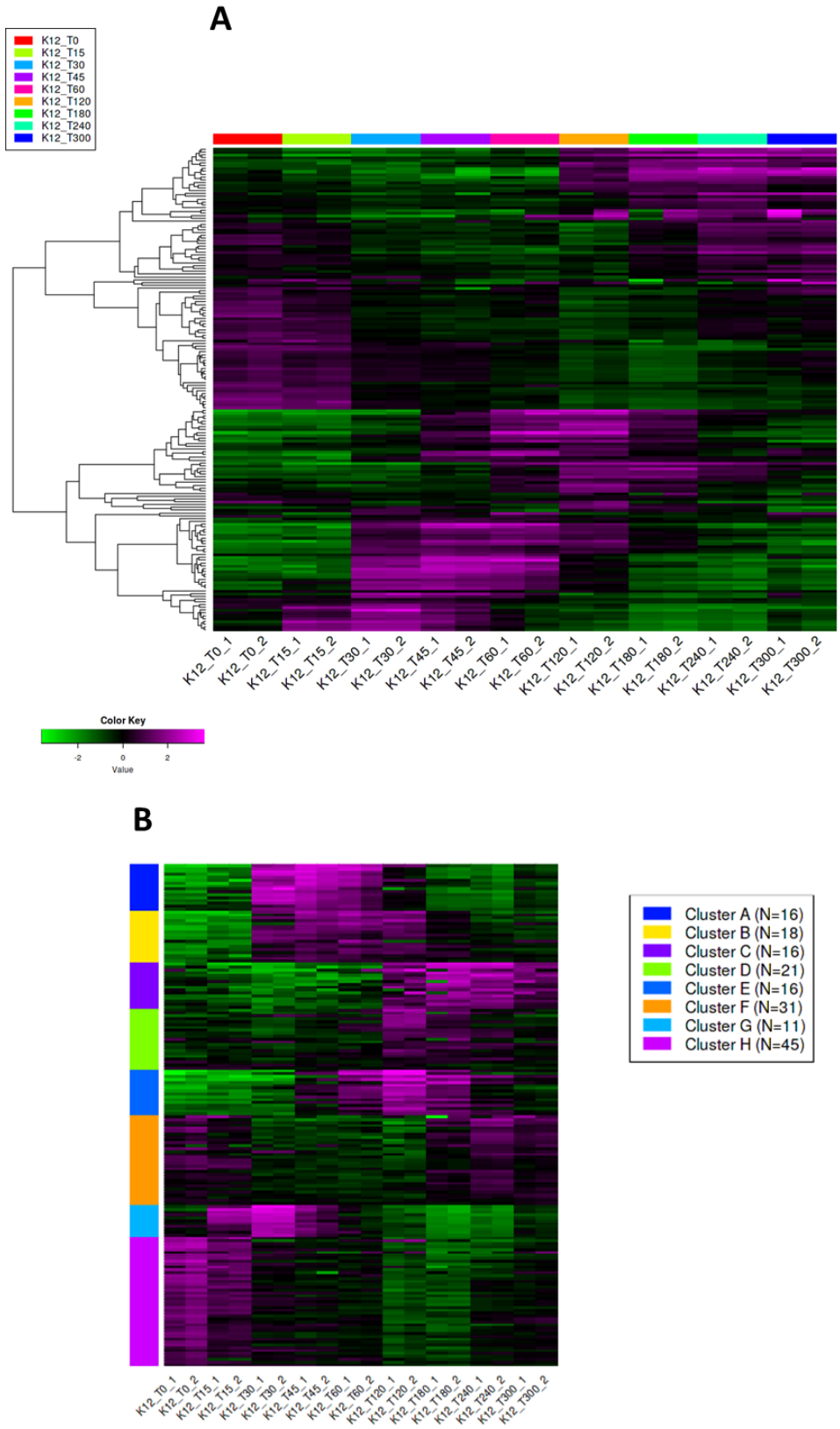

Figure 35 A- dendrogram of hierarchical clustering for genes from Experiment 1 that were upregulated in response to both biofilms and nutrients in Experiments 5 and 6 **B-** *k*-means clustering to 8 clusters for these genes.

Figure 35 shows results of combining significantly differentially regulated genes in both response to biofilms (Experiment 5) and nutrients (Experiment 6) and examining their expression patterns throughout the predation cycle on *E. coli* (Experiment 1). Taking all genes differentially regulated (called as $q<1 \times 10^{-5}$ in Rockhopper) in both response to nutrients and biofilms resulted in 246 genes which are expressed at various points throughout the predation cycle and this immediately validates the authors' suggestion that there are many responses in common to the two conditions.

Cluster E consists of genes upregulated throughout the intraperiplasmic growth cycle and these clusters support the authors' conclusions that there is a general response to nutrients (both from broth or released from dispersing biofilm), with Cluster E enriched for pathways of transcription, translation and protein and nitrogen metabolism.

Clusters A and C are genes related to attack phase, with those in Cluster C highly upregulated at time 0 and 240 minutes (at release of new attack phase cells) and enriched for flagellar motility and pili genes. Those in Cluster A are expressed at 300 minutes and also have many pilus-associated genes.

Those of Cluster B are upregulated upon contact with prey and include the operon *bd2224-bd2229* which encodes a putative transmembrane import/export system.

Repeating this analysis with genes that were either exclusively called as differentially regulated in response to nutrients but not biofilms and vice versa, surprisingly gave similar results to the above. Pathways for transcription, translation and protein and nitrogen metabolism were enriched in both cases, as they were for the dataset that overlapped between the two conditions. Similarly, the surprising upregulation of genes that are upregulated upon contact with prey were found in both datasets, with different L,D-transpeptidases upregulated in different datasets. Closer inspection revealed that many were from the same operons (e.g. the operon *bd0412-bd0420* had representatives in the overlapping dataset as well as each non-overlapping dataset) and so it is likely that virtually all of these are essentially the same response, but some genes were scored as significant only in one dataset, whilst some scored significantly in both. Thus attempts to discover any clear differences between the two responses were unsuccessful, suggesting that the responses are very similar.

Conclusions from Experiment 6

Experiment 6 shows a strong, global transcriptional response to nutrients with a large set of genes upregulated to metabolise the nutrients and initiate growth and cell elongation, but not cell division. The response to nutrients was very similar to the response to *S. aureus* biofilms where again, the *B. bacteriovorus* respond in order to utilise any available nutrients in both conditions. What is surprising is that some genes which are very specific to predation of Gram-negative bacteria, such as prey cell wall modifying enzymes are also upregulated despite there not being a use for these in the conditions in which they are being upregulated. This suggests that something within both nutrient broth and *S. aureus* biofilms includes the signal that *B. bacteriovorus* take from their Gram-negative prey to induce expression of their arsenal of predation genes alongside general metabolism genes. This highlights the specialist predator nature of *B. bacteriovorus* in that they do not appear to have a general nutrient scavenging strategy, only a switch on of predation genes or not.

Experiment 7- *B. bacteriovorus* response to Diffusible Signal Factor

Diffusible Signal Factor (DSF) is a quorum sensing molecule that is toxic to *B. bacteriovorus*. To determine the effect of this toxicity, global transcription was compared between attack phase cells and intraperiplasmic *B. bacteriovorus* with and without DSF

in buffer. This experiment was carried out on the strain *B. bacteriovorus* 109J. Strain 109J is >90% identical to HD100, so it was possible to map this transcriptome to the different strain, with 401 genes and 533 RNAs being called as significantly upregulated (q<1 x 10^{-5}) in the presence of DSF, however only 14 tRNAs and one antisense RNA were in common with the data from Experiment 1, so this analysis couldn't add significantly to the conclusions of the paper.

Experiment 8- Prey interactions of different predation-deficient host independent *B. bacteriovorus* mutants

Three different predation-deficient host-independent mutants were used in this experiment; Δ*mglA* mutants are capable only of transient attachment to prey, Δ*bd1291* form an irreversible attachment to prey, but fail to enter the prey and Δ*bd2473* mostly form an irreversible attachment with only a small percentage successfully entering the prey. None of the mutants are capable of predatory growth and could only be rescued by host independent growth. In order to determine which early prey interaction genes are deficient in these mutants, they and control strains HID13 and HID22 were resuspended in Ca/HEPES buffer and presented with prey *E. coli* for 15 minutes and compared to controls in Ca/HEPES buffer only.

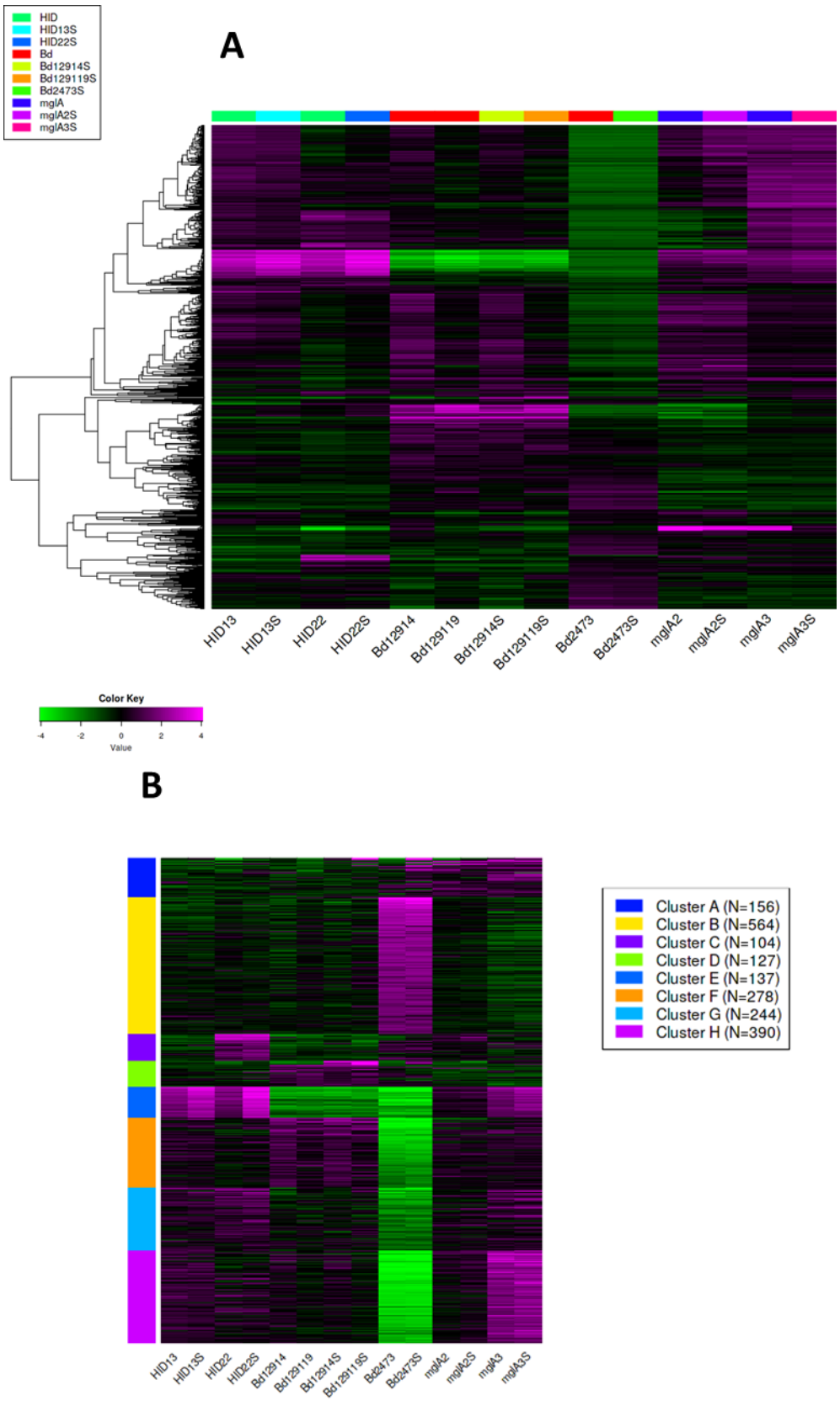

Figure 36 A- dendrogram of hierarchical clustering for genes from Experiment 8 **B-** *k*-means clustering to 8 clusters for these genes. HID13 and HID22 are wild-type predation-competent controls. +S; HI strain presented with *E. coli* S17-1

Figure 36 shows the results of clustering for the various host-independent mutants when presented to prey *E. coli* S17-1 (+S) and controls in buffer only. The first conclusion is that gene expression is very strain dependent in HI strains, with large clusters of genes expressed both with and without prey for some strains and not for others. This is particularly marked for the Δ*bd2473* strain, which had a drastically different expression pattern compared with the other strains. Similarly, even strains with the same gene deleted had vastly different expression patterns, for example strains 2 and 3 of Δ*mglA* had significant differences with a large set of genes upregulated in strain 3 in Cluster H, but not in strain 2.

Cluster A consists of genes that are upregulated in both Δ*mglA* strains (with generally higher expression in strain 3 than strain 2). This cluster includes the operon *bd2224-bd2229* encoding a putative transmembrane transport system. The cluster also includes siderophores and the gene *bd1291* itself.

Cluster B consists of genes highly expressed in the *Δbd2473* strain, but not in any of the other strains and likely represents the global differences in expression between HI strains. Amongst a wide variety of genes in this cluster is *mglA.* The genes in this cluster are highly expressed in the Δ*bd2473* strain regardless of mixing with prey.

Similarly, Cluster C consists of genes highly expressed in the HID22 strain relative to the other strains, again representing global differences between the different HI strains.

Cluster D consists of genes upregulated in both of the Δ*bd1291* strains. Whilst it is possible that these are also a result of HI variation, as most of the genes in this cluster are upregulated in both strains, then it is probable that these are as a result of the directed gene deletion. This cluster contains some flagellar motility genes and several non-coding RNAs.

Cluster E is the most useful for comparison between the mutants as this consists of genes upregulated in contact with prey for the wild-type strains HID13 and HID22, but with a much reduced effect if any in the *ΔmglA* strains and highly repressed in both *Δbd1291* strains and in the Δ*bd2473* strain, suggesting that these are the genes involved in prey interaction in which the mutant strains are defective at correctly regulating. This cluster includes the prey wall modifying enzymes such as *bd0993* which encodes a peptidoglycan deacetylase, the L,D-transpeptidases *bd1176, bd1358* and *bd3176*, carboxypeptidases *bd0816* and *bd3459,* lytic transglycosylase *bd3575* and lysozyme *bd1411.* This cluster also contains the operon *bd0412-bd0420* which has homology to gliding motility genes and is upregulated on prey contact (in Experiment 1).

Clusters F, G and H consist mostly of genes strongly suppressed in the Δ*bd2473* strain, with varying responses in the other strains. These are most likely formed as a result of the HI specific variation and notably contain a lot of flagellar motility and chemotaxis genes, suggesting that this strain may be defective in swimming.

Conclusions from Experiment 8

Experiment 8 shows that there is drastically different expression patterns for different HI strains. In this experiment, the HI cells were grown to mid-log phase (OD_{600} 0.6) in rich broth media, then spun down and resuspended in Ca/HEPES buffer. Any growth of HI cells results in a mix of morphologies and stages of growth for the cells and this varies drastically from strain to strain and is a result of the different mutations which have allowed the strains to grow host independently. It is likely this variation that is reflected in the varied expression patterns seen between the strains. This effect was seen most in the Δ*bd2473* strain and it may be for this strain, a combination of the differences in HI strains and also an effect of the targeted deletion resulted in such different gene

expression patterns. One possibility is that RNA from this preparation was different to the other preparations as a result of differing cell morphology/growth stages, but this seems unlikely as the QC suggested the same quality levels. Unfortunately, there was only one strain isolated (as it may be difficult to isolate such a defective strain), so it wasn't possible to compare multiple different HI strains for this mutant as it was for the other strains. Bd2473 is predicted to be a TsaP protein which stabilises the outer membrane PilQ porin through which the pilus passes. The *B. bacteriovorus* protein is predicted to have an extended C-terminus which may extend through the periplasm to an inner membrane domain. Our hypothesis is that this protein, in addition to stabilising PilQ, may also have a signalling role, detecting the pilus status and affecting global gene regulation in response to this status. This is supported by the observation that the majority of attacks on prey abort at the stage of attachment and the observation here of a globally different expression pattern compared to the other strains further supports this idea. It is particularly notable that the genes from Cluster E, which are upregulated upon contact with prey in the wild-type control (many of which are known to be involved in prey modification) are strongly repressed in this strain, even in the presence of prey.

The two Δ*mglA* strains also showed some considerable variation between them, further supporting the conclusion that the HI phenotype itself is responsible for a large degree of variation in gene expression. MglA organises the pilus at the attacking pole of *B. bacteriovorus* and its absence renders *B. bacteriovorus* incapable of irreversible prey attachment with fewer pili observed at the pole. The Δ*mglA* strain only undergoes reversible attachment to prey, detaching again a few minutes. This is reflected in the transcription in Experiment 8 as none of the genes upregulated upon interaction with prey in the wild-type controls (in Cluster E) were upregulated in the Δ*mglA* strains. Unlike the Δ*bd2473* and Δ*bd1291* strains, however, rather than being repressed, these genes were expressed at similar levels to those of the uninduced (no prey added) control wild-type strains, suggesting that the Δ*mglA* mutation in itself does not drastically alter gene expression, rather it simply does not allow the *B. bacteriovorus* to attach correctly to prey for gene induction to occur.

In contrast, Cluster D had a group of genes that were common to both strains of Δ*bd2191*, suggesting that these may be differentially regulated as a result of the directed mutation rather than general HI strain variation. Bd1291 is annotated as PilG, an accessory gene associated with pili of unknown function. The gene *bd1291* appears to be in an operon with the major pilin protein PilA, which has been determined to be the main pilin which forms pili in *B. bacteriovorus* and is essential for prey entry (Evans et al., 2007, Capeness et al., 2013). The predicted protein Bd1291 also has a TPR domain predicted to be involved in protein-protein interactions for pili signalling. Deletion of *bd1291* results in the *B. bacteriovorus* cell forming an irreversible attachment to the prey cell, but with predation stalled at this step and no entry into the prey cell. As is for the case with the Δ*bd2473* strain, the hypothesis is that Bd1291 is involved in sensing the pilus attachment/retraction status and signalling to enable prey entry. Cluster E shows that genes involved in early prey predatory interaction, upregulated in the wild-type strains, are strongly repressed in the Δ*bd1291* strain, supporting the idea that this gene product is important in global predatory gene regulation.

Conclusions

Here, I have developed a robust pipeline for data extraction, trimming, quality control, mapping, counting and clustering for the analysis of RNA-Seq data from experiments with *B. bacteriovorus*. Using these methods, new insights were gleaned from data that

had not yet been fully analysed, or in other groups' data that had been analysed in different ways.

Firstly, the precise temporal regulation of clusters of genes at distinct times throughout the predatory lifecycle of predation on *E. coli* was shown. For some of these, clues as to the functions of these was shown by annotated genes, for example genes associated with RNA degradation and metabolism were amongst the first group of genes to be upregulated (then downregulated shortly after, having completed their function). For others groups, further bioinformatics analysis was needed to predict their function, many of which had previously been the focus of work in our lab. Amongst the most interesting of these were the genes sharply upregulated upon prey contact and these include some degradative enzymes such as nucleases and proteases and genes encoding cell wall modification enzymes, some potentially for modifying the predator's own wall, many others for modifying the prey wall. The cluster analysis allows us to group other genes of unknown function with these genes known to be important for predation to give indicators for future studies. The cluster analysis of the predation on *E. coli* (Experiment 1) gives researchers an invaluable tool for studying genes upregulated at all stages of the predation cycle. It also serves as a reference dataset to compare other RNA-Seq experiments with to determine which groups of genes attributed to different stages of predation in this experiment were expressed in different conditions.

Secondly, cluster analysis showed that predation by *B. bacteriovorus* on *Serratia* was not synchronous relative to the timepoints of predation on *E. coli*, with groups of genes that were associated with different stages of predation on *E. coli* expressed throughout the different timepoints in this experiment (Experiment 2). The same was true of the mutant strain Δ*dgcC*, which was defective in the guanidine cyclase gene responsible for regulating the switch to HI growth. Analysis revealed differences in global regulation between this mutant and the wild-type, with large groups of genes (particularly associated with flagellar motility) expressed at different levels compared to wild-type.

Initial analyses of *B. bacteriovorus* response to both nutrients and *S. aureus* biofilms had suggested that this was similar to intraperiplasmic growth (Im et al., 2018). Cluster analysis reveals that surprisingly, the response includes the upregulation of genes specifically involved in Gram-negative prey modification, suggesting that utilisation of nutrients is intrinsically linked to predation in this obligate predator and that upregulation of genes involved in the former globally induces upregulation of genes involved in the latter.

The analyses comparing predation in, and exposure to, pooled human serum revealed that the response of *B. bacteriovorus* seemed to be one of protection against the antibacterial elements of serum, with toxin antiporters, outer membrane components and iron-associated products such as siderophores produced. Despite serum potentially being a nutrient rich environment, the response serum was not similar to the response to nutrients, which could bode well for the potential of *B. bacteriovorus* as a therapeutic.

Analyses of HI mutant strains defective at different stages of predation showed both that individual HI strains have drastically different global expression profiles and that there are clusters of genes likely associated with the specific mutations tested, giving insight into how these mutations are disrupting the predation process.

Along with insights into the predation process and condition responses, the gene clusters generated in this project identify many targets for future projects to better understand the predation process and how *B. bacteriovorus* reacts in more clinically relevant conditions; a prerequisite for the fulfilment of its promise as a potential novel antimicrobial therapy.

References

ATTERBURY, R. J., HOBLEY, L., TILL, R., LAMBERT, C., CAPENESS, M. J., LERNER, T. R., FENTON, A. K., BARROW, P. & SOCKETT, R. E. 2011. Effects of orally administered *Bdellovibrio bacteriovorus* on the well-being and *Salmonella* colonization of young chicks. *Applied and Environmental Microbiology,* 77**,** 5794-803.

BAREL, G. & JURKEVITCH, E. 2001. Analysis of phenotypic diversity among host-independent mutants of *Bdellovibrio bacteriovorus* 109J. *Archives of Microbiology,* 176**,** 211-6.

CAPENESS, M. J., LAMBERT, C., LOVERING, A. L., TILL, R., UCHIDA, K., CHAUDHURI, R., ALDERWICK, L. J., LEE, D. J., SWARBRECK, D., LIDDELL, S., AIZAWA, S. & SOCKETT, R. E. 2013. Activity of *Bdellovibrio hit* locus proteins, Bd0108 and Bd0109, links Type IVa pilus extrusion/retraction status to prey-independent growth signalling. *PLoS One,* 8**,** e79759.

DASHIFF, A., JUNKA, R. A., LIBERA, M. & KADOURI, D. E. 2011. Predation of human pathogens by the predatory bacteria *Micavibrio aeruginosavorus* and *Bdellovibrio bacteriovorus*. *Journal of Applied Microbiology,* 110**,** 431-44.

DASHIFF, A. & KADOURI, D. E. 2011. Predation of oral pathogens by *Bdellovibrio bacteriovorus* 109J. *Molecular Oral Microbiology* 26**,** 19-34.

EVANS, K. J., LAMBERT, C. & SOCKETT, R. E. 2007. Predation by *Bdellovibrio bacteriovorus* HD100 requires type IV pili. *Journal of Bacteriology,* 189**,** 4850-9.

FRATAMICO, P. M. & WHITING, R. C. 1995. Ability of *Bdellovibrio bacteriovorus* 109J to lyse gram-negative food-borne pathogenic and spoilage bacteria. *Journal of food Protection,* 58**,** 160-164.

GE, S. X., SON, E. W. & YAO, R. 2018. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics,* 19**,** 534.

HARDING, C. J., HUWILER, S. G., SOMERS, H., LAMBERT, C., RAY, L. J., TILL, R., TAYLOR, G., MOYNIHAN, P. J., SOCKETT, R. E. & LOVERING, A. L. 2020. A lysozyme with altered substrate specificity facilitates prey cell exit by the periplasmic predator *Bdellovibrio bacteriovorus*. *Nat Commun,* 11**,** 4817.

HESPELL, R. B., MIOZZARI, G. F. & RITTENBERG, S. C. 1975. Ribonucleic acid destruction and synthesis during intraperiplasmic growth of *Bdellovibrio bacteriovorus*. *Journal of Bacteriology,* 123**,** 481-91.

HOBLEY, L., FUNG, R. K., LAMBERT, C., HARRIS, M. A., DABHI, J. M., KING, S. S., BASFORD, S. M., UCHIDA, K., TILL, R., AHMAD, R., AIZAWA, S., GOMELSKY, M. & SOCKETT, R. E. 2012a. Discrete cyclic di-GMP-dependent control of bacterial predation versus axenic growth in *Bdellovibrio bacteriovorus*. *PLoS Pathog,* 8**,** e1002493.

HOBLEY, L., LERNER, T. R., WILLIAMS, L. E., LAMBERT, C., TILL, R., MILNER, D. S., BASFORD, S. M., CAPENESS, M. J., FENTON, A. K., ATTERBURY, R. J., HARRIS, M. A. & SOCKETT, R. E. 2012b. Genome analysis of a simultaneously predatory and prey-independent, novel *Bdellovibrio bacteriovorus* from the River Tiber, supports in silico predictions of both ancient and recent lateral gene transfer from diverse bacteria. *BMC Genomics,* 13**,** 670.

IM, H., DWIDAR, M. & MITCHELL, R. J. 2018. *Bdellovibrio bacteriovorus* HD100, a predator of Gram-negative bacteria, benefits energetically from *Staphylococcus aureus* biofilms without predation. *ISME J,* 12**,** 2090-2095.

JURKEVITCH, E., MINZ, D., RAMATI, B. & BAREL, G. 2000. Prey range characterization, ribotyping, and diversity of soil and rhizosphere *Bdellovibrio* spp. isolated on phytopathogenic bacteria. *Applied and Environmental Microbiology,* 66**,** 2365-71.

KADOURI, D. E., TO, K., SHANKS, R. M. & DOI, Y. 2013. Predatory bacteria: a potential ally against multidrug-resistant Gram-negative pathogens. *PLoS One,* 8**,** e63397.
KURU, E., LAMBERT, C., RITTICHIER, J., TILL, R., DUCRET, A., DEROUAUX, A., GRAY, J., BIBOY, J., VOLLMER, W., VAN NIEUWENHZE, M. S., BRUN, Y. V. & SOCKETT, E. R. 2017. Fluorescent D-amino-acids reveal bi-cellular cell wall modifications important for *Bdellovibrio bacteriovorus* predation including L,D-transpeptidase mediated prey strengthening. *Nature Microbiology,* in press.
LAMBERT, C., CHANG, C. Y., CAPENESS, M. J. & SOCKETT, R. E. 2010a. The first bite-profiling the predatosome in the bacterial pathogen *Bdellovibrio*. *PLoS One,* 5**,** e8599.
LAMBERT, C., EVANS, K. J., TILL, R., HOBLEY, L., CAPENESS, M., RENDULIC, S., SCHUSTER, S. C., AIZAWA, S. & SOCKETT, R. E. 2006. Characterizing the flagellar filament and the role of motility in bacterial prey-penetration by *Bdellovibrio bacteriovorus*. *Molecular Microbiology,* 60**,** 274-86.
LAMBERT, C., FENTON, A. K., HOBLEY, L. & SOCKETT, R. E. 2011. Predatory *Bdellovibrio* bacteria use gliding motility to scout for prey on surfaces. *J Bacteriol,* 193**,** 3139-41.
LAMBERT, C., HOBLEY, L., CHANG, C. Y., FENTON, A., CAPENESS, M. & SOCKETT, L. 2009. A predatory patchwork: membrane and surface structures of *Bdellovibrio bacteriovorus*. *Advances in Microbial Physiology,* 54**,** 313-61.
LAMBERT, C., IVANOV, P. & SOCKETT, L. 2010b. A Transcriptional "Scream" Early Response of *E. coli* Prey to Predatory Invasion by *Bdellovibrio Current Microbiology* 60**,** 419-427.
LAMBERT, C. & SOCKETT, R. E. 2013. Nucleases in *Bdellovibrio bacteriovorus* contribute towards efficient self-biofilm formation and eradication of preformed prey biofilms. *FEMS Microbiol Lett,* 340**,** 109-16.
LAMBERT, C., TILL, R., HOBLEY, L. & SOCKETT, R. E. 2012. Mutagenesis of RpoE-like sigma factor genes in *Bdellovibrio* reveals differential control of groEL and two groES genes. *BMC Microbiol,* 12**,** 99.
LEONARDY, S., MIERTZSCHKE, M., BULYHA, I., SPERLING, E., WITTINGHOFER, A. & SØGAARD-ANDERSEN, L. 2010. Regulation of dynamic polarity switching in bacteria by a Ras-like G-protein and its cognate GAP. *The EMBO Journal,* 29**,** 2276-2289.
LERNER, T. R., LOVERING, A. L., BUI, N. K., UCHIDA, K., AIZAWA, S., VOLLMER, W. & SOCKETT, R. E. 2012. Specialized peptidoglycan hydrolases sculpt the intra-bacterial niche of predatory *Bdellovibrio* and increase population fitness. *PLoS Pathogens,* 8**,** e1002524.
LIVERMORE, D. M. 2009. Has the era of untreatable infections arrived? *J Antimicrob Chemother,* 64 Suppl 1**,** i29-36.
MATIN, A. & RITTENBERG, S. C. 1972. Kinetics of deoxyribonucleic acid destruction and synthesis during growth of *Bdellovibrio bacteriovorus* strain 109D on *Pseudomonas putida* and *Escherichia coli*. *Journal of Bacteriology,* 111**,** 664-73.
MCCLURE, R., BALASUBRAMANIAN, D., SUN, Y., BOBROVSKYY, M., SUMBY, P., GENCO, C. A., VANDERPOOL, C. K. & TJADEN, B. 2013. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res,* 41**,** e140.
MILNER, D. S., TILL, R., CADBY, I., LOVERING, A. L., BASFORD, S. M., SAXON, E. B., LIDDELL, S., WILLIAMS, L. E. & SOCKETT, R. E. 2014. Ras GTPase-Like Protein MglA, a Controller of Bacterial Social-Motility in Myxobacteria, Has Evolved to Control Bacterial Predation by *Bdellovibrio*. *PLoS Genet,* 10**,** e1004253.
NEGUS, D., MOORE, C., BAKER, M., RAGHUNATHAN, D., TYSON, J. & SOCKETT, R. E. 2017. Predator Versus Pathogen: How Does Predatory *Bdellovibrio bacteriovorus* Interface with the Challenges of Killing Gram-Negative Pathogens in a Host Setting? *Annu Rev Microbiol,* 71**,** 441-457.
PAREKH, S., ZIEGENHAIN, C., VIETH, B., ENARD, W. & HELLMANN, I. 2016. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep,* 6**,** 25533.

REINER, A. M. & SHILO, M. 1969. Host-independent growth of *Bdellovibrio bacteriovorus* in microbial extracts. *Journal of General Microbiology,* 59**,** 401-410.
RENDULIC, S., JAGTAP, P., ROSINUS, A., EPPINGER, M., BAAR, C., LANZ, C., KELLER, H., LAMBERT, C., EVANS, K. J., GOESMANN, A., MEYER, F., SOCKETT, R. E. & SCHUSTER, S. C. 2004. A predator unmasked: life cycle of *Bdellovibrio bacteriovorus* from a genomic perspective. *Science,* 303**,** 689-92.
SHATZKES, K., CHAE, R., TANG, C., RAMIREZ, G. C., MUKHERJEE, S., TSENOVA, L., CONNELL, N. D. & KADOURI, D. E. 2015. Examining the safety of respiratory and intravenous inoculation of *Bdellovibrio bacteriovorus* and *Micavibrio aeruginosavorus* in a mouse model. *Sci Rep,* 5**,** 12899.
SHATZKES, K., SINGLETON, E., TANG, C., ZUENA, M., SHUKLA, S., GUPTA, S., DHARANI, S., ONYILE, O., RINAGGIO, J., CONNELL, N. D. & KADOURI, D. E. 2016. Predatory Bacteria Attenuate *Klebsiella pneumoniae* Burden in Rat Lungs. *MBio,* 7.
SHEMESH, Y. & JURKEVITCH, E. 2004. Plastic phenotypic resistance to predation by *Bdellovibrio* and like organisms in bacterial prey. *Environ Microbiol,* 6**,** 12-8.
STOLP, H. & PETZOLD, H. 1962. Untersuchungen uber einen obligat parasitischen mikroorganismus mit lyticher aktivitat fur *Pseudomonas*-bakterien. *Phytopathologische Zeitschrift,* 45**,** 364-390.
STOLP, H. & STARR, M. P. 1963. *Bdellovibrio bacteriovorus* gen. et sp. n., a predatory, ectoparasitic, and bacteriolytic microorganism. *Antonie van Leeuwenhoek Journal of Microbiology and Seriology,* 29**,** 217-248.
THOMASHOW, M. F. & RITTENBERG, S. C. 1979. Descriptive biology of the bdellovibrios. *In:* PARISH, J. H. (ed.) *Developmental biology of prokaryotes.* 9th ed.: University of California Press.
VARON, M. & SHILO, M. 1969. Attachment of *Bdellovibrio bacteriovorus* to cell wall mutants of *Salmonella* spp. and *Escherichia coli*. *Journal of Bacteriology,* 97**,** 977-9.
WILLIAMS, C. R., BACCARELLA, A., PARRISH, J. Z. & KIM, C. C. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics,* 17**,** 103.
WILLIS, A. R., MOORE, C., MAZON-MOYA, M., KROKOWSKI, S., LAMBERT, C., TILL, R., MOSTOWY, S. & SOCKETT, R. E. 2016. Injections of Predatory Bacteria Work Alongside Host Immune Cells to Treat *Shigella* Infection in Zebrafish Larvae. *Curr Biol,* 26**,** 3343-3351.
WURTZEL, O., DORI-BACHASH, M., PIETROKOVSKI, S., JURKEVITCH, E. & SOREK, R. 2010. Mutation detection with next-generation resequencing through a mediator genome. *PLoS One,* 5**,** e15628.

Appendix

A- Script for merging datasets

Python script "pandas_merge_inner.py" for merging common data from two .csv files. This merges the data that each file has in common in the "Synonym" column, which represents the Bdxxxx gene number. As predicted RNAs did not have this, the Synonym column for these was replaced by the starting position of the RNA.

Script:

```
import pandas as pd


# read csv files
```

```
data1 = pd.read_csv('input_filename_1.csv')
data2 = pd.read_csv('input_filename_2.csv')

# merge files
output1 = pd.merge(data1, data2,
                   on='Synonym',
                   how='inner')

# show result
print(output1)

# write to csv
output1.to_csv('Output_filename.csv', header=True)
```

B- *K*-means clustering datafile information

All final *k*-means clustering data used in this study (i.e. excluding datasets which were used as exploratory to hone the pipeline) are in the Excel spreadsheet "Final *K*-means clustering data.xls". Each tab is named by the figure in which the data is presented and row 1 of each tab contains a description of the data within.