2015

Machine Learning on Images of a Microbial Mutant Library

Georgia Doing
Bard College

Machine Learning on Images of a Microbial Mutant Library

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
of
Bard College

by
Georgia Doing

Annandale-on-Hudson, New York
May, 2015

Abstract

Environmental isolates, like BJB312, are interesting because of their potential therapeutic properties. Transposon mutagenesis is a technique used to determine the function of genes by randomly disrupting a genome and observing the phenotypic effects. The genome of BJB312 consists of over 5,000 genes, requiring 57,000 independent insertion mutants in order to break every gene in the genome. It is unwieldy to screen such a large library for defects. I used image processing techniques to convert qualitative data of mutant bacterial colonies morphology into a quantitative data set that is susceptible to data mining. Further, I built a tool of ensemble machine learning techniques that automatically analyze a large library of mutants. It first uses the unsupervised methods k-means and Wards hierarchical clustering to find a patterned, recurrent phenotype. It then uses a Support Vector Machine to screen the library at large. This tool is robust and useful on real-world data because it utilizes Machine Learning techniques to filter the image library before reaching the final clustering solution. Ten transposon insertion mutants that clustered together were characterized by lessened biofilm. This proof-of-concept study shows that genomic and high-throughput functional characterizations can be combined in order to rapidly explore a novel microbe.

Contents

List of Figures

Dedication

This is for A.H.Y. As is all I do.

Acknowledgments

With infinite thanks for my advisors Dr. Brooke Jude and Dr. Rebecca Thomas
who have taught me to love looking for answers
and helped me daily in that uphill battle;
with thanks for Craig, Rebecca, Dwayne and Maureen
who helped me solve so many problems, including the centrifuge I broke;
with thanks for Susan Fox Rogers
who taught me how to write for joy;
with thanks for Dr. Mike Tibbetts and Dr. Keith O'Hara
who helped me realize, when I was a first year, that I liked science;
with thanks for all the professors
who made my time at Bard one of my most fulfilling;
with thanks for my dear friends Dylan and Elisa
who science is lucky to have on its side;
with thanks for Rylan
who has walked with me over the past four years;
with thanks for Melissa
who spent hours with me in the lab, witnessing my blunders;
with thanks for the Microbiology class of Spring '15
who built half of the BJB312 mutant library;
with thanks for my Brother
who is my brother;
with thanks for my Dad
who is my dad;
with thanks for my Mom
who is my mom;

many thanks for what you've done.

1
Introduction

Large mutant libraries are powerful tools for characterizing a novel strain of bacteria. Libraries can be screened for phenotypes and the genes associated with that trait can be determined. Libraries are generally made as large as possible in order to increase the chances of disrupting more genes and gaining a fuller annotation of the genome. However, as libraries increase in size it becomes unwieldy to screen them in fully functional assays. Consequently, mutant libraries are made but typically only a couple of the mutant strains are examined for their phenotype and so function is only assigned to a few genes at a time based on the abnormal trait displayed.

I have made a tool that allows for the comprehensive exploration of a mutant library and a functional assessment of its genome by automating functional assay screening (Figure 1.0.1). Image processing techniques allow qualitative data of microbial phenotypes to be translated into a large quantitative data set. This transformation is crucial because it makes the library susceptible to data mining. Machine Learning techniques such as clustering and support vector machines can analyze the large, heterogeneous data set and find patterns that would not be practical to find by hand.

This tool does not require any knowledge of the mutant library. It uses unsupervised machine learning to find recurrent patterns of phenotypes. Subsequently, a phenotype of interest can be explored to saturation using supervised SVM classification; once a phenotype group has been identified, all close or related phenotype mutants can be collected from the mutant library and all relevant genes can be identified. The ensemble of these unsupervised and supervised machine learning methods creates a complete workflow that can discover new phenotypes and generate training data and then use that training data to characterize the entire library with respect to a phenotype of interest.

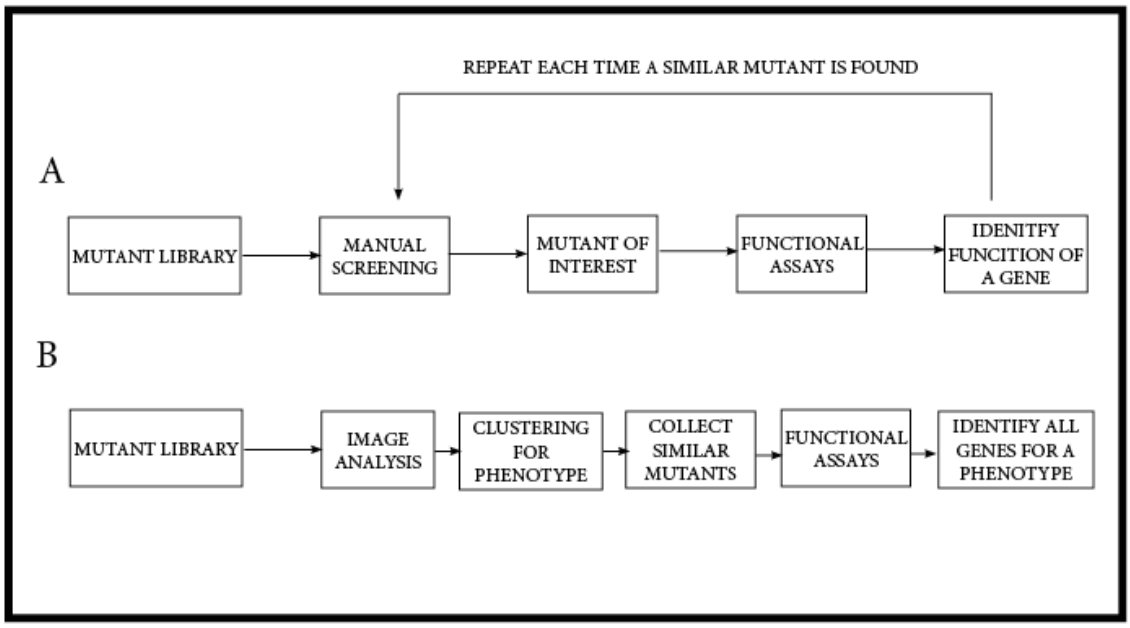

Figure 1.0.1. (A) The standard workflow for exploring a library of mutants needs to be repeated in order to identify all genes that contribute to a phenotype. (B) The method developed in this study could potentially explore a phenotype to saturation in one iteration.

This tool is robust because it uses Machine Learning to filter the data set before performing clustering and classification. Practically speaking, large, manually-plated mutant libraries will have some errant elements. Image processing tools first address this by only capturing mutants that can be identified using a Hough Circle Transform. Secondly,

k-means sequesters repeating artifacts. Thirdly, a one-class SVM refines the data set so that it only includes mutants that appear morphologically different from the wild type. These filtering methods improve the statistical measures and biologically usefulness of clustering solutions.

In this study, I characterized a library of $\sim$ 4,000 BJB312 mutants and found a group of biofilm-deficient mutants. This study does not fully characterize BJB312 or its biofilm characteristics, but these scale vertically and horizontally; they could be applied to a larger library in order to reach saturation of a phenotype; they could be used to analyze a library repeatedly form different perspectives, investigating a variety of phenotypes.

2
Background in Biology

2.1 Culturing Microbes in the Laboratory.

The mechanisms of microbes can be untangled when the microbes are cultured in a laboratory. In a controlled laboratory setting, microbial strains are cultured and subjected to genetic, molecular, biochemical and physiological tests so that their characteristics can be better known. Isolated environmental strains of bacteria can be studied in controlled conditions to reveal their individual functionalities. Isolating bacteria allows for the study of these characteristics. Many environmental strains have effects on human health and the ecosystem, and environmental microbiological surveys allow for the determination of potential health and environmental risks [Guo et al., 2015, Landell et al., 2013]. It is important to study bacterial strains that cause disease such as *Pseudomonas aeruginosa*, *Escherichia coli* and *Staphylococcus aureus* in order to develop effective treatments [Ruiz et al., 2004, Kosmidis et al., 2012, Titilawo et al., 2015] . By looking closely at clinical and environmental isolates, we can begin to understand these bacterias strengths and weaknesses in order to combat disease, propel industries, and study environments.

Determining how microbes resist antibiotics or infect humans allows for the development of technologies that combat these things. Culture assays are the primary ways to determine how microbes work, which are important because of the ubiquitous direct and indirect interactions we have with microbes in our shared environment.

2.2 The Hudson Valley Watershed

Bacterial environmental isolates can provide information about the ecosystem, and have the potential to be used as therapeutics. Water samples from different fresh water bodies throughout the Hudson Valley Watershed can provide samples of the environmental bacteria present (Figure 2.2.1). By isolating bacteria from different areas of the Hudson Valley Watershed and culturing them to figure out their traits, they can be compared. Surveys of environmental isolates can reveal new bacteria that have unique and useful attributes.

Various violacein-producing bacteria have been isolated from fresh water samples of the Hudson Valley Watershed (Figure 2.2.2). Some of these strains have been identified as *Janthinobacterium*. *Janthinobacterium* are Gram-negative bacteria found in aqueous and soil environments. They grow at room temperature, but the full range of temperatures they can live in is not specifically defined. They characteristically display antibiotic resistance form biofilms and produce metabolites such as hydrogen [Ning et al., 2012, Pidot et al., 2013, Xia et al., 2014]. In particular, a species of *Janthinobacterium* that is known to produce violacein is *Janthinobacterium lividum*. Violacein producers are important to study in the laboratory because they have potential therapeutic properties.

2.3 Violacein

Violacein is a purple pigment indole derivative of tryptophan [Duran et al., 2007]. It is produced by some well-studied microbes such as *Chromobacterium violacein* and *Jan-*

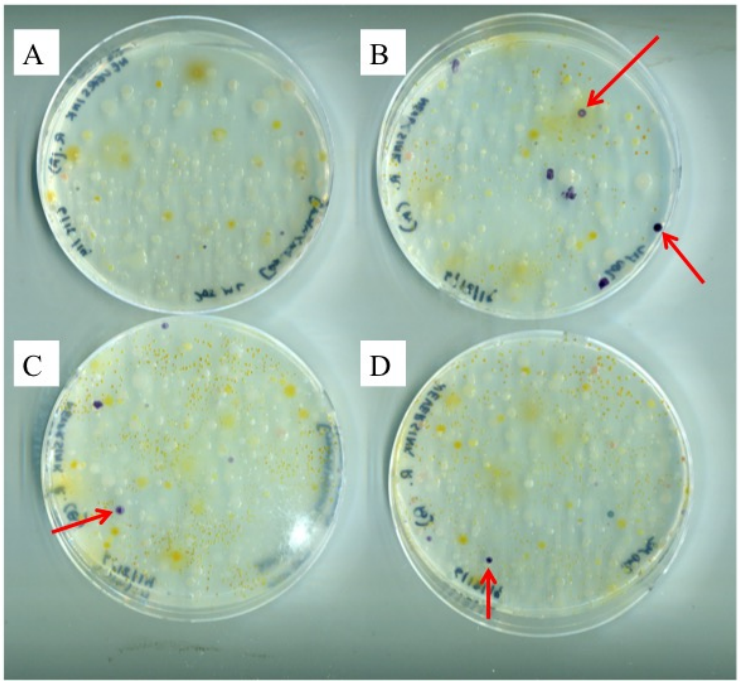

Figure 2.2.1. Environmental microbes from the Hudson Valley Watershed growing on R2A minimal media > 48 hours. Violacein producers appear with various phenotypes and shades of purple (arrows).

thinobacterium lividum as well as a plethora of unstudied environmental isolates [Duran and Menck, 2001, Pantanella et al., 2006, Hoshino, 2011]. There is a breadth of therapeutic properties associated with this purple compound. There are indications that it is protective for many reasons. For example, studies suggest it is anti-parasitic, anti-viral, and anti-fungal [Lopes et al., 2009, Andrighetti-Frohner et al., 2003, Andrighetti-Fröhner et al., 2006]. It has even been considered for its activity against cancers, including human leukemia [Kodach et al., 2006, Ferreira, 2004]. In corroboration with the utility of violacein, studies have shown that violacein is produced in response to stress. Physical agitation of bacterial colonies caused aggregation and violacein production [Yang et al., 2007]. Given these properties, it is evident that producing violacein during times of duress is advantageous.

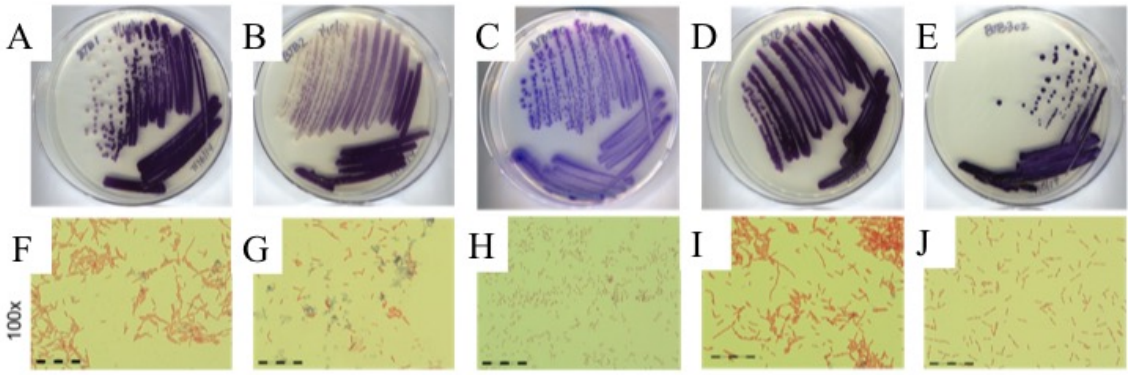

Figure 2.2.2. (A-E) Pure cultures of environmentally isolates violacein producers from the Hudson Valley fresh water samples on (A-E) LB and (C) on R2A). (F-J) Gram stains of each strain show colony morphology and indicate Gram negative bacteria. Scale bar = $5.58\mu m$.

What is known about violacein comes from studies of well-known strains in the laboratory and can inform the exploration of new isolates. The operon responsible for violacein production, the *vio* operon, was discovered through studying *Ch. violaceum* [August et al., 2000]. The *vio* operon consists of five genes: *vioA, vioB, vioC, vioD, vioE* (Figure 2.3.1). The genes textitvioA-vioE work together to convert tryptophan into violacein Duran:2007de. Other bacteria, such a *J. lividum*, have since been shown to contain a homologous operon in their genome.

Beyond violacein production, the effectiveness of violacein and extent to which the properties of the pure chemical are practically advantageous depend on the program of

violacein production, violacein release, and ultimately survival of the violacein-producers. These, in turn, depend on biofilm.

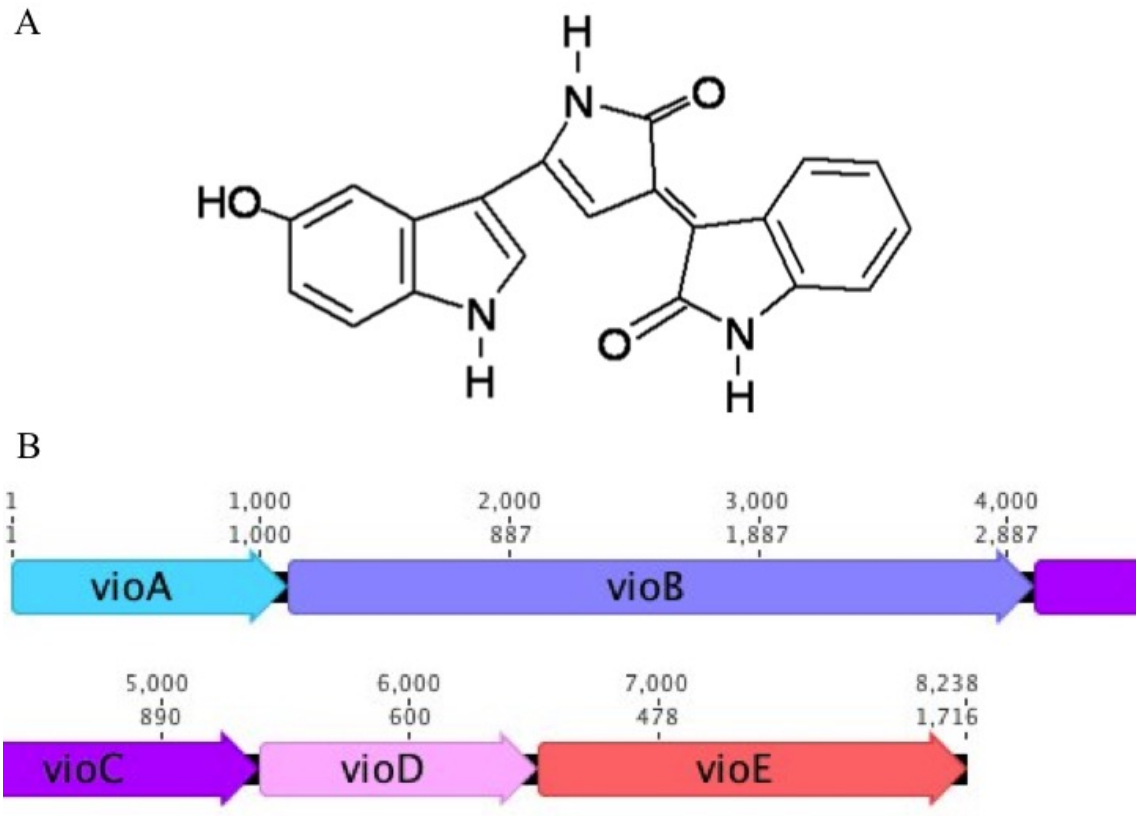

Figure 2.3.1. (A) The chemical structure of violacein. (B) The five gene *v*io operon from *J. lividum*, responsible for violacein production.

2.4 Biofilm

Biofilms are organized communities of microbes encapsulated in extracellular polymeric substances growing in functional structures, containing nucleic acids proteins, minerals, nutrients and cell wall material to promote bacterial growth, and increasing their fitness and survival [Costerton et al., 1987, Carpentier and Cerf, 1993]. The stability of the bacteria allow for architectures to be built and maintained. The survival of many microbes, especially those in aqueous environments, often depends on the robustness of their

biofilm; taller, more differentiated, more rugose biofilms are often associated with large aggregates of cells whereas smooth, while short and uniform biofilms are more susceptible to detergents like sodium dodecyl sulfate [Davies et al., 1998]. This example illustrates the importance of biofilm and how it promotes microbial communities that are very different than motile, suspended ones.

Although the bacterial cells are covered in extracellular material, the architecture of biofilms is highly porous and, as many biofilms are located in aqueous environments, is open to a constant flow of water [Lawrence et al., 1989]. Acting as a filter, the physical structure of biofilms is thus advantageous because it protects cells from threats of the external environment such as predation or antibiotics, while selectively allowing open exchange with desired parts of the environment, such as nutrients, from the water [Carpentier and Cerf, 1993, Coquet et al., , Mah and O'Toole, 2001]. For example, a type of architecture that roughly resembles a mushroom architecture provides these intricate services. Two common architectural elements seen in *P. aeruginosa* biofilms are stalks and caps that together form mushroom-like structures; stalk cells are attached to the surface are form the base of the biofilm and hold up cap cells, which project further into the environment [Pamp et al., 2008]. Bacterial stalk cells are morphologically distinct from the cap cells, have lower metabolisms and are non-motile. On the other hand, cap cells have higher metabolisms, and are more like aggregates of motile bacteria [Pamp et al., 2008, Alavi and Belas, 2001].This pattern can be advantageous, allowing nutrients to diffuse more easily into the peripheral parts of the biofilm. The structural differences in the stalk and cap bacteria are demonstrated by the differential effects the antibiotics tetracycline and colistin have on cap and stalk bacteria respectively [Haagensen et al., 2006, Bjarnsholt et al., 2005].

Biofilms are protective and have been observed to form when bacteria are in adverse situations, such as under the stress of starvation, antibiotics and foreign immune

responses [Landry et al., 2006]. Biofilms show more resistance to antibiotics, antibodies, surfactants, bacteriophages and predators like amoeba than planktonic bacteria [Hentzer et al., 2001]. For example, to increase its tolerance to the antibiotic tobramycin biofilm becomes mucoid by overproducing alginate [Stapper, 2004, Hentzer et al., 2001]. Because biofilms are associated with relatively stable population sizes, low growth, they are able to avoid the activity of many bacteriostatic antibiotics [Eng et al., 1991]. The thickness of biofilms affects the metabolic rate of bacteria in the community, thus affecting the rate at which cells require nutrients and take in exogenous molecules; when cells require fewer nutrients, having an antibiotic deprive them of nutrients has less of an impact [Whitford and Schumacher, 1964].

Biofilm formation is a highly organized process. Both the temporal and spatial regulation of biofilm are important; when microbes identify a favorable environment, they grow in abundance until they are able to expend energy on functions other than rapid replication, setting up permanent shop. Biofilms begin to be seen in their stationary phase and are maximal in the early death stages. Biofilm formation also depends on their surrounding, on whether a favorable carbon source, such as glycerol, is present and on whether it is faced with stressors such as ampicillin [Pantanella et al., 2006]. Biofilms are also affected by factors such as pH, oxygen levels, carbon sources, osmolarity, and friction levels during growth (affecting attachment and thickness) [O'Toole et al., 2000, Biggs and hickey, 1994, Graba et al., 2013].

Biofilm begins with primary adhesion, or stochastic events of contact between planktonic organisms and a conditioned surface [Marshall et al., 1971]. This form of attachment is reversible and allows bacteria to sample surfaces while determining a good location for adhesion. Beyond observing a potential surface, bacteria condition surfaces to make conditions more favorable for interactions, for example, by producing biosurfactants such as rhamnolipids [Tremblay et al., 2007, Gristina, 1987]. While making primary adhesions and

with the encouragement of biosurfactants, bacterial cells can perform a different, surface based motility called swarming and form a monolayer, migrating around the surface before aggregating [O'Toole and Kolter, 1998]. For example *P. aeruginosa* has been reported to use twitching motility and chemotaxis to move up gradients of phospholipids, suggesting like activities could be employed when exploring new surfaces [Kearns et al., 2001]. After a suitable surface has been sampled, bacterial cells make a permanent attachments to the surface. In this process, adhesion proteins make molecular and specific binds with a conditioned surface, anchoring the bacteria to the surface. The complexity of biofilm formation illustrates the breadth of genes that play a role in biofilm regulation; biofilm is an exemplary polygenic complex trait. This, it is important to study it using high-throughput genetic studies.

There are a number of genes identified as important genes in biofilm formation in *J*anthinobacterium species, including *p*el and *p*sl, which code for exopolysaccharides involved in attachment and encasing [Friedman and Kolter, 2004, Jackson et al., 2004, Matsukawa and Greenberg, 2004]. In addition, the *r*hlI and *r*hlR pathways are important because rhamnolipids are crucial for maintaining the pores in biofilms by disruption cell-cell and cell-substrate connections and encouraging swarming [Davey et al., 2003]. This pathway is controlled by QS, specifically c-di-GMP binds to *p*elD and increases EPS production [Lee et al., 2007]. This concurs with data suggesting a link between QS and biofilm production [Davies et al., 1998]. Adhesions are transcriptionally regulated and can be turned on depending on the environment [Ziebuhr et al., 1999].

2.5 Chytridiomycosis

The survival, violacein production and biofilm formation of violacein producers, like *J*anthinobacterium, are important because of the ecological role they play, defending amphibians from the fungal pathogen *Batrachochytrium dendrobatidis* (*B*d) (Figure 2.5.1).

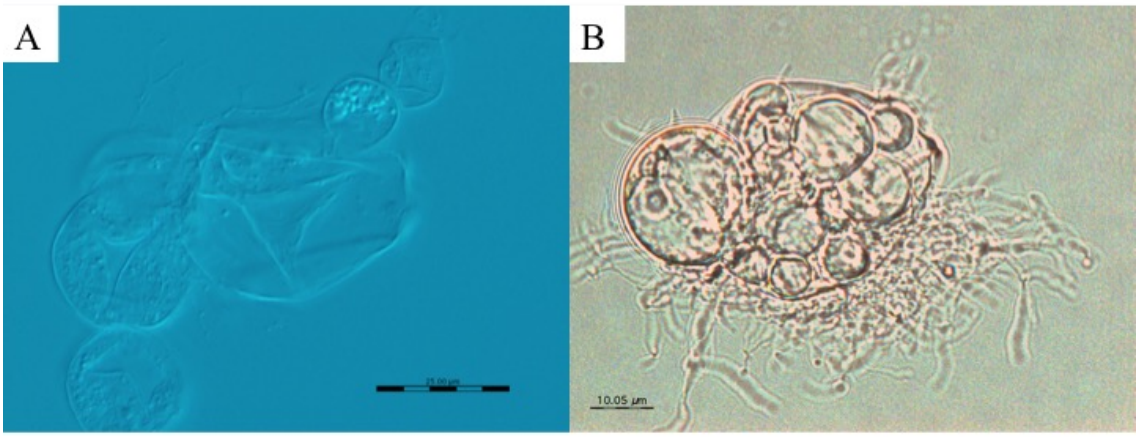

Figure 2.5.1. Images of *Batrachochytriumum dendrobatidis*, strain JEL423, under (A) DIC microscopy,scale bar = 25μm, and (B) bright field microscopy, scale bar = 10.05μm.

Bd is a fungal pathogen that causes a lethal skin disease called chytridiomycosis (chytrid) in amphibian populations worldwide (Figure 2.5.2). As of 2014, 32.5% of all amphibian species were threatened by Bd [Longcore et al., 2014]. Some symptoms of chytrid are tissue erosion, hyperkeratosis, hyperplasia, weight loss, and death [Berger et al.,]. Chytrid may have been responsible for declines in amphibian population since the 1970s, and is becoming more prolific [Corn and Fogleman, 1984, Carey, 1993].

As with the decline of any species, the devastation of amphibians means more than just the loss of amphibian populations. Loss of any population corresponds to a loss of local diversity, which has many ramifications [Ouellet et al., 2005, **?**]. The food web in which the amphibians participate loses balance as the resources they consumed become

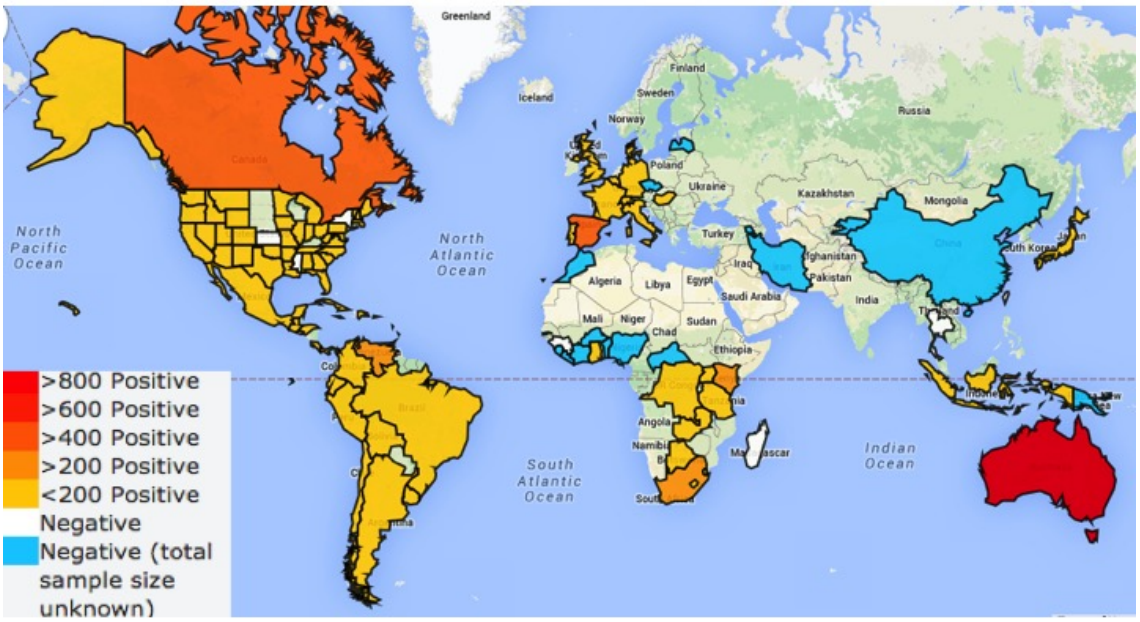

Figure 2.5.2. World map indicating the current prevalence of Bd from www.Bd.net accessed in March of 2015. Inset key colors indicates the number of reported cases.

overabundant and those they produced become scarce. Diversity loss is associated with the increased spread of disease effecting members across a community [Keesing et al., 2010]. Humans are part of the ecological systems in which diseases are spreading more easily, and it behooves us to understand the mechanisms to counter *B*d. *B*d is the only chytrid within the rhizophydiales clade to infect vertebrates [Joneson et al., 2011]. On vertebrates, *B*d lives on skin and strictly infects keratinocytes [Symonds et al., 2008]. Although amphibians are known to excrete anti-microbial peptides onto their skin as an immune response, this mechanism doesnt appear to be acting to counter *B*d [Rollins-Smith et al., 2002, Pessier et al., 1999]. Perhaps the pathogen is successful because, being a chytrid, it is unlike other

amphibian pathogens and does not trigger the release of defense peptides, circumventing the system that protects frogs from other pathogens.

2.6 Bioaugmentation and *J*anthinobacterium

Amphibian populations are not able to defend themselves from *B*d, but some have potentially partnered up with protective microbes. Studies have reported mutualism between cutaneous bacteria and amphibian hosts such as the eastern red-backed salamander, female four-toes salamanders and mountain yellow-legged frogs, *Rana muscosa* [Harris et al., 2009, Lauer et al., 2007a, Woodhams et al., 2011]. Further surveys have observed that amphibian populations that sustainably coexists with *B*d have higher levels of protective bacteria than similar amphibian populations that are experiencing rapid Bd induced decline [Woodhams et al., 2007, Lam et al., 2010]. This suggests that bacteria act to subdue *B*d, allowing the cohabitation of *B*d and amphibians without as rapid death. Some of these bacteria are violacein-producers (Figure 2.6.1).

Protective bacteria can fight off pathogens by producing anti-fungal, anti-microbial, or otherwise therapeutic metabolites. Violacein has been shown to repel and inhibit *B*d growth in the laboratory (data not shown). This is not surprising given the reported properties of violacein: antibacterial, antiviral and anti-tumoral among others [Lopes et al., 2009, Andrighetti-Frohner et al., 2003, Becker and Harris, 2010, Brucker et al., 2008]. Field studies have shown that, when comparing amphibians that both have microbes including *J. lividum*, those with higher amounts of violacein on their skin survive better [Brucker et al., 2008]. The chemical nature of violacein is interesting, given its therapeutic properties. The pathway from tryptophan to violacein shares similarities with the anabolism of antimicrobials such as the DNA binding inhibitor rebeccamycin and the protein kinase inhibitor staurospirine [Moreau et al., 1999, Tamaoki et al., 1986]. Given these properties, it is not surprising that friendly bacteria have been shown to repel Bd, reducing

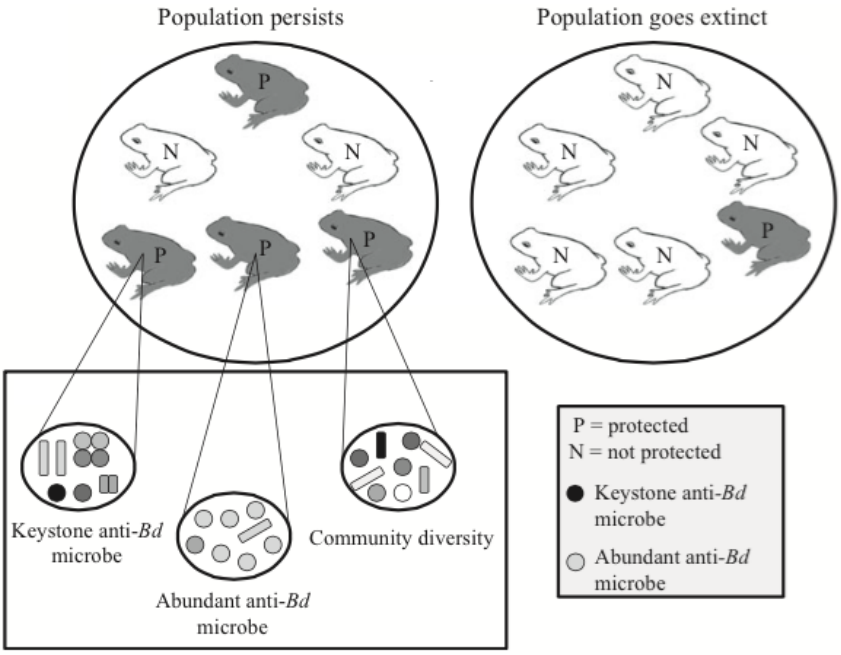

Figure 2.6.1. Potential ecological interactions between amphibians, Bd and violacein-producing bacteria [Bletz et al., 2013].

the need for an immune response in amphibians [Lam et al., 2011](Lam, Walton, and Harris). Additionally, protective bacteria help their host by enhancing the hosts innate immune system [Lauer et al., 2007a, Meyer et al., 2012]. This type of defense against Bd is highly favorable because it helps amphibians protect themselves instead of attempting to artificially remove the threat.

Conservation methods have explored different methods of fighting chytrid. *B*d zoospores are sensitive to fungicides and extreme heat [Johnson and Speare, 2003]. However, many treatments such as these, that are successful in a laboratory, are not realistically applicable to natural ecosystems. Conservation efforts must recognize natural dynamics between species in order to be effective. Well-deigned bioaugmentation recognizes relationships between organisms in an environment and capitalizes on pre-existing mutualism. Specifically, bioaugmentation efforts could (and have began to) harness the

power of violacein-producing cutaneous bacteria that live on amphibians to protect their hosts from *B*d.

Bioaugmentation, when designed well, is a superior conservation method because it can work with the normal flora and fauna and be self-propelling. Firstly, probiotics can reproduce and sustain themselves as a population Bletz:2013ka. Because bacteria often use metabolites to fend off pathogens, it stands to reason that the species of bacteria is not as important for effective protection as much as genes that code for a specific metabolite. Considering that the *v*io operon is near a movable element in genomes such as *J. lividum*, it is possible that genetic transfer can help proliferate protective bacteria beyond just the colonization of a single species. Protective qualities can transfer through a microbial community vertically, horizontally or environmentally. These methods may be advantageous because they provide the opportunity for protective bacteria to become prevalent enough to be effective without a microbiome turning into a monoculture; a diverse range of organisms can be protective. This will help heed the principle of disrupting natural balance as little as possible and thus avoiding the adverse effects such as increase disease spread that accompany loss of diversity [Belden and Harris, 2007]. Preliminary laboratory experiments have demonstrated that bacteria, such as *J. lividum*, have reduced the effects of *B*d [Lauer et al., 2007b]. Although these reports bode well for using bioaugmentation in the field there remain many unaccounted variables. Efforts have tried species-specific baths [Harris et al., 2009]. Further, community-based treatments in water or soil inoculation have been used to protect salamanders with J lividum [Muletz et al., 2012]. When implementing bioaugmentation, careful consideration must be taken when deciding which probiotic to introduce to an environment. Universally, a good probiotic should be a good competitor on ventral surfaces, limbs and feet in order to co-habitate with *B*d [North and Alford, 2008]. Further, it should make metabolites, like violacein, at low density or grow to high density quickly. A similar consideration is that it should proficiently produce

violacein because violacein is only effective past certain concentration thresholds [Becker et al., 2009]. Biofilm formation is an important factor because it plays a role in adhering to amphibian skin and survival and is associated with violacein production and dispersal. Different probiotics will be better for different environments depending on their metabolism; *I*odobacter are anaerobic and *J. lividum* is aerobic, and *I*odobacter and *Ch. violaceum* grow at different temperatures. The least disruptive bioaugmentation agent would be the one native to a given amphibian population.

2.7 Environmental Isolates

Ultimately, it is crucial to explore environmental violacein-producers like BJB312 while deciding which bacterial species could be used in different populations in order to best respect natural dynamics, avoid side-effects, and provide the most effective treatment. Many different bacteria produce violacein including *M*assilia, *J. lividum*, *D*unganella and *C*ollimonas [Yada et al., 2008, Agematu et al., 2011, Aranda et al., 2011, Hakvåg et al., 2009]. The breadth of violacein producers illustrates the potential of bioaugmentation techniques but also calls for closer characterization of each microbe in order to use them appropriately and with the highest yield. These violacein producers have been found in soils and sea-waters across the world in diverse countries such as Japan, China, Spain and Norway. In order to design effective treatments, first we need to know how violacein-producers work. By culturing violacein producers and exploring their genome to study violacein production properties and biofilm formation we can better apply bioaugmentation to treat *B*d.

2.8 Genomics

Genome data for environmental isolates is more readily available because of advances in next generation sequencing technology. Sequences for entire bacterial genomes

are affordable and provide a wealth of information. Software such as BLAST (Basic Local Alignment Search Tool) and Geneious make comparative genomics simple across national databases like PubMed and novel, in-lab studies.

Illumina sequencing is a next generation sequencing technique that allows for the rapid sequencing of entire microbial genomes by using parallel pyrosequencing. Sequence reads can be assembled together based on overlapping regions with tools such as Geneious, NextGene or MIRA. These assembly tools use de Brujin graph theory to efficiently search and match variable length sequences. The appropriate parameters such as k-mer length, can be determined using the Velvet Optimizer. Once the whole genome is assembled, annotation tools, like RAST (Rapid Annotation using Subsystem Technology), can match regions of the genome with known genes. By comparing genetic sequences, the nearest neighbors of an isolate can be identified and can potentially be placed in an operational taxonomic unit. Furthermore, knowing the sequence of an organism allows for targeted mutagenesis and in-frame gene deletion. Before such experiments can be truly useful, the isolate should be explored functionally. Having an isolates genome aids in initial un-targeted exploration, as annotated genes are categorized into subsystems such as metabolism, resistance, and secretion systems. Thus, the genetic composition of an organism can suggest traits and inform functional assay design.

2.9 Functional Assays

In order to compliment genomic data and learn about the functional traits of new strains of bacteria, novel environmental isolates can be subjected to transposon mutagenesis. Transposon mutagenesis is a technique that can be used in conjunction with culturable assays. If a strain of bacteria is culturable in the laboratory, it can be genetically manipulated to reveal the functions of its genes. Unlike gene deletion, this is a useful technique because an organism can be studied from the point of view of a phenotype of interest, such

as biofilm, before the genes responsible for the phenotype can be identified, allowing for the identification of novels phenotypes and mechanisms. For example I study profile bacteria based on their biofilm morphology, which is not a trait that could be identified by a operon in a genome [Queiroz et al., 2012]. In transposon insertion mutagenesis, the genome of an organism is perturbed by the incorporation of an exogenous fragment of DNA. The function of the broken gene can be inferred by the resulting loss-of-function phenotype. Insertion mutagenesis can be site-directed, in which case it reveals the phenotype of a specific gene of interest, or it can be random and allow for the genetic cause of an observed phenotype to be discovered. Both methods rely on the logic of using loss-of-function phenotype to infer gene function.

Random transposon mutagenesis allows for the undirected exploration of a genome. In random mutagenesis, the transposon is designed to invade the hosts genome at an unspecified loci, allowing any gene to be disrupted. This technique allows for the discovery of novel mechanisms because it allows for the disruption of genes that werent previously known to exist. It is important to be able to uncover these mechanisms; being able to understand a pathogenic bacteria would allow us to identify and design a potential treatment [Autret and Charbit, 2005]. Most transposon mutagenesis techniques are limited to the recovery of insertions in only non-essential genes because they rely on the growth of the mutant for isolation. However, some methods, like TnAraOut, have found ways to recover insertions in essential genes [Judson and Mekalanos, 2000].

Transposon mutagenesis has been used to identify many important genetic pathways. For example, the *v*irB operon which is responsible for virulence factors and lipopolysaccharides in the Gram-negative human pathogen *Brucella suis* was identified with transposon mutagenesis [Foulongne et al., 2000]. Genes involved in colonization by *Vibrio cholera* have also been identified using transposons [Chiang and Mekalanos, 1998]. Transposon mutagenesis has been performed in many microbes including *P. aeruginosa*, *Methanosarcina*

acetiverans, *Mycobacterium tuberculosis* and *M*ycoplasma [Pelicic et al., 1997, Hutchison III, 1999]. Because it has been successfully performed in a wide range of organisms, it is a good candidate tool to use on environmental isolates and other previously unexplored strains.

To accompany the breadth of studies that use transposon mutagenesis, there are may different designed for transposon vectors. One such plasmid, built by Rachel Larson called pRL27, contains five critical pieces in a Tn5 backbone: an origin of transfer (oriT), a pi-dependent origin of replication (oriR), kanamycin resistance, and two primers (Figure 2.9.1). Although these five elements are essential, there is variation in the versions of plasmids; different plasmid designs use different marker and selection elements such as a *l*acZ gene or tetracycline resistance [Metcalf et al., 1996,Larsen et al., 2002,de Lorenzo and Timmis, 1994]. In transposon insertion mutagenesis it is often preferred to use a suicide vector, like that in the Tn5 transposon, which cannot replicate independently in their host and thus can only be incorporated into a hosts genome once.

Because it controls for a single insertion even, pRL27 is a good candidate plasmid for building a single insertion transposon mutant library that contains mutants with disruptions of every gene in the genome. Such a highly saturated library contains a comprehensive set functional data and can be paired with genomic sequencing to map the functions of every gene in the genome. Because this type of mutagenesis is random, it cannot be used to systematically target each gene in a genome. However, probability theory can be used to estimate the number of mutants a library must contain in order for it to likely contain a disruption of every gene in a genome. When disruption in every gene is achieved, and a genome has been saturated, the resulting library contains enough information to build a comprehensive functional profile. It is not straightforward to accurately estimate the number of mutants required to build a saturation library for any given genome. Some of the factors include overall base-pair length of the genome, the number of genes, the bp

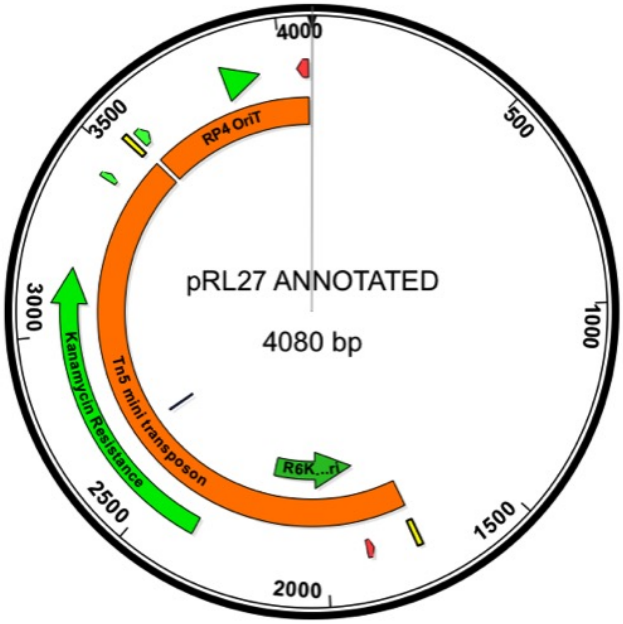

Figure 2.9.1. Plasmid map of pRL27 annotated with a Kanamycin resistance cassette, origin of transfer, origin of replication, transposable element and outward facing primers [Larsen et al., 2002]. Graphics made in Geneious.

lengths of genes, the number of vital genes, and the bias of hotspots [Lefevre and Watkins, 1986]. Many studies determine their library size empirically, by conducting increasingly large scale mutagenesis followed by large scale sequencing in order to determine the number of novel mutants and when that number cease to increase [Phogot et al., 2001, Jacobs et al., 2011] (PHOGAT et al.; Jacobs et al.). This number depends so heavily on the bacteria strain of focus and the method of mutagenesis. However, since not all studies have the resources to empirically determine an appropriate library size, many rely on estimates. Statistical models that include many of these variables have been used to predict saturation levels, but there is always a degree of estimation to a model as well as degrees of uncertainty inherent in unknowns for which the model cannot account. The Poisson distribution has provided a foundational formula for estimations of acceptable sizes for saturation libraries PHOGAT:2001wc.In addition, the Gamma distribution is a candidate

for such a model because it can account for the variable mutation rates across different genes (Figure 2.9.2) [Pollock and Larkin, 2004]. There is not yet a standard model amongst these candidates.

Number of events (plants) required to be screened at 0.95 probability*			
Number of targets (x)	Using $f(i)^{\#} = e^{-m}m^i/i!$, where $m = r/n$	Using random hit generator$^{\spadesuit}$	Using $f(i)^{\clubsuit} = e^{-m}m^i/i!$, where $m = ne^{-r/n}$
5	15	21	23
10	30	50	53
15	45	84	86
25	75	155	155
50	150	339	345
75	225	543	547
100	300	751	758
125	375	970	975
250	750	2131	2123
500	1500	L$^{\spadesuit}$	4593
1000	3000	L	9878
2000	6000	L	21,143
5000	15,000	L	57,437
10000	30,000	L	1,21,806
20000	60,000	L	2,57,474
40000	1,20,000	L	5,46,674

Figure 2.9.2. Estimations of the number of insertion events required to saturate genomes with 4 though 40,000 genes according to two models and experimental results from a simulation. [Phogot et al., 2001]

2.10 High Throughput Functional Assays

If the sequence of an isolates genome is available, inferences can be made about the isolate's phenotypic properties, such as metabolic bi-products, antibiotic resistance and carbon utilizations can be made by comparing the sequences of an environmental isolate to sequences of standard strains and well-studied genomes. However, inferences must be supported by experimental data when studying a trait in depth. Further, the function of novel

genes cannot be assumed from comparable genomes, it must be discovered experimentally. In order to keep up with the pace of information technology and the accumulation of genomic data laboratory protocols, like transposon mutagenesis, high-throughput demands must be accommodated. The practicality of a high-throughput mutagenesis depends on the genome and transposon, but thousands of mutants can often be produced in a day or two. However, it is not as easy to collect and screen such large libraries. Collection protocols and assays need to be adjusted to isolate and collect phenotypic information from mutants quickly. This is the essential challenging in trying to gather functional data at the same pace as genetic data can be acquired.

When high-throughput functional analyses are able to produce enough mutants, genomes can be understood at a rate that keeps up with genomic analysis. Genomes diverse as *E. coli*, *P. aeruginosa*, *Burrelia burgdophia*, *Beauveria bassiana* fungus *D*rosophila and *M*aize have been explored in this way [Tong et al., 2004, Lefevre and Watkins, 1986, Walbot, 2000, Fan et al., 2011, Jacobs et al., 2011, Stewart et al., 2004]. These researchers aimed for saturation; they tried to break every gene in the genome at least once by processing enough mutants that success was statistically likely. When a mutant library contains a mutant with a gene disruption even for every gene in a genome, that genome has become saturated. As long as functional data from these mutants is collected and analyzed, saturation mutagenesis gives a comprehensive profile a whole the genome.

3
Background in Computer Science

3.1 Image Processing

There is a wealth of image data available from sources such as the internet, and images are relatively easy to capture. Sites like Facebook and Instagram have made taking and sharing images comparably as common as sharing textual information. Digital cameras, technologies attached to microscopes, and even cell phones make gathering image data trivial. In particular, they have allowed for the gathering of visual information from high throughput, genetic, microbial functional studies.

The field of computer vision, or image processing, aims to extract meaningful information from digital images. The digital image is generally represented by a matrix of 1 to 4 channels: grayscale, red-green-blue or hue-saturation-value with an optional alpha channel. Typically, the range of the channel values is 0-255 where low values are dark and high values are bright. The enormous availability of image data has made the fields of computer vision and image processing popular. There are many techniques and tools that can be used to process images and extract information about their composition and/or content from their raw pixel values (without meta data).

3.1.1 Pre-processing

Images are often pre-processed so they are optimally susceptible to image processing techniques. There are some standard processing techniques that can reduce the noise of an image and make the salient features easier to detect. One way to do this is using linear filters to transform one image into another, thereby capturing or calculating a certain quality of the image. Two common types of linear transformation are correlation and convolution. Both are functions, of an NxN mask, F, and an image, I (Eq. 3.1.1 and Eq. 3.1.2).

$$F * I(x, y) = \sum_{j=-N}^{N} \sum_{i=-N}^{N} F(i, j)I(x - i, y - j) \tag{3.1.1}$$

$$F * I(x, y) = \sum_{j=N}^{N} \sum_{i=N}^{N} F(i, j)I(x + i, y + j) \tag{3.1.2}$$

In both of these procedures, a smaller 'mask' image is run across a larger image, reassigning the image's pixel value to the sum or average of the image's pixels multiplied by its corresponding mask pixels. Both processes require edge case specifications because the pixels that are near an edge cannot be computed in the standard way. Edge cases can be handled in any manner of ways for example by using half the kernel, copying the most adjacent pixel values, leaving the edges unchanged or padding the image with zeros so that the edge cases use a black border. Which technique that is best depends on the level of accuracy needed in the resulting image and the importance of the edges.

Because each pixel of the resulting image depends on the values of pixels that are in front of and behind it, these processes cannot be done in place. Further, these processes have time complexity proportional to the number of pixels in an image. However, the complexity of the procedures can be reduced by either separating a 3x3 kernel into two

linear 3-element kernel. Therefore, although the time complexity is potentially high, these techniques are still convenient to use on large sets of data.

Blurring the image by convolving it with a blur kernel reduces minute, noisy, variations in pixel values. Two standard kernels are the box and Gaussian kernels. The box kernel averages a pixel with its neighbors (Eq. 3.1.3). The Gaussian kernel also averages a pixel with its neighbors but is more sophisticated because it weighs nearer neighbors more heavily than more distant ones (Eq. 3.1.4).

$$S_{box} = \frac{1}{16}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3.1.3}$$

$$S_{gaussian} = \frac{1}{256}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} \tag{3.1.4}$$

The optimal size and shape of the blurring kernel depends on the individual image data and the scale of the target features that will eventually be extracted. If an image has high resolution and the features of interest are relatively coarse, the image can be blurred more dramatically and made easier to process. On the other hand, if the image is low resolution or a pattern of interest is very fine, too much blurring can obscure the content of the image.

Just as blurring an image can make it easier to process, transforming an image into an edge image can also be convenient. Two more useful linear filters are the Sobel, which detects edges and a corner detector . The Sobel kernel detects differences between adjacent pixels (Eq. 3.1.5). It can be oriented different ways to detect edges that run in different directions. Corners are slightly more advanced features to use for image transform, but can still be captured with linear filtering (Eq. 3.1.6).

$$S_{sobel} = \frac{1}{8}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{3.1.5}$$

$$S_{corner} = \frac{1}{4}\begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} \tag{3.1.6}$$

Making an image binary can make it easier to process because the set of pixel values is reduced to {0,255}, in other words a lot of noise has been removed from the signal. For features such as edges, shapes, contours and objects the relevant information is the absence or presence of a feature in the image, not its qualities. Therefore, a binary representation suffices. An image can be made binary using a threshold function, θ, such that any pixel, p, with a value above the threshold, t, will be reassigned to 255 and any pixel with a value below the threshold will be assigned to 0 (Eq 3.1.7).

$$\theta(p,t) = \begin{cases} 1 & p > t \\ 0 & else \end{cases} \tag{3.1.7}$$

Using variations on binary images, for example pixels above the threshold keep their original value, images can be made into tool such as masks that remove background. The biggest challenge in transforming an image to binary is deciding on the threshold value. There may be scenarios when the threshold is easily decided from the content of the images, but often the goal is to have a binary image that maintains as much contour and shape information from the original image as possible and no other parameters are known. For this, the optimal threshold depends on the composition of an image in particular. There are algorithms to determine an optimal threshold. For example, the Otsu method uses descriptive statistics of an image's histogram to choose an optimal threshold (Eq. 3.1.8). The method aims to find two distributions within the histogram, q_1, q_2, one representing lower values and one representing higher values. It is designed to work best on bimodal images. It chooses a threshold, t that minimizes the variance, σ_1^2, σ_2^2, within each half of

the image [Szeliski, 2010, Beghdadi et al., 2013]. In this way, the threshold can be tailored to each image upon which it is imposed.

$$\sigma_w^2 = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \tag{3.1.8}$$

Blurring can remove noise, but it also lessens the sharpness of an image. Often, it is desirable to remove noise form an image while keeping it sharp. A combination of erosion and dilation can work to remove specks of noise while sharpening boarders, significant edges and contours at the same time. Erosion and dilation are reciprocal procedures. Both convolve the image, f, with a kernel, s where S is the size of s, and can be represented by thresholding functions, σ. In erosion, the kernel calculates the minimum value of all the pixels within reach (Eq. 3.1.9). In dilation the kernel calculates the maximum value of all the pixels in its kernel region (Eq. 3.1.10).

$$erode(f, s) = \theta(c, S) \tag{3.1.9}$$

$$dilate(f, s) = \theta(c, 1) \tag{3.1.10}$$

As consequences of these operations, erosion can shrink the size of an image and dilation can grow it. However, if the operations are used in succession, the image can end up relatively the same size as its original by either shrinking the object then growing (opening) it or vice versa (closing) (Eq 3.1.11 and Eq. 3.1.12).

$$open(f, s) = dilate(erode(f, s), s) \tag{3.1.11}$$

$$close(f, s) = erode(dilate(f, s), s) \tag{3.1.12}$$

In a standard black and white image, for example, wherein the background is black and an object in the image is white, erosion will eliminate noise by counting any noisy signal and background, and dilation will remove noise by including any noisy signal as part of the object (Eq. 3.1.1).

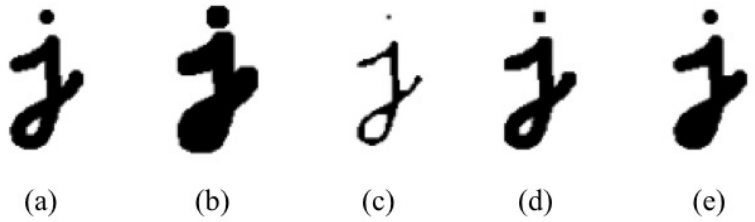

Figure 3.1.1. Morphological transforms can maintain certain aspects of an image while removing some of the signal to make it easier to work with. (a) original image, (b) dilation, (c) erosion, (d) opening, (e) closing.

With real world data, it is often necessary to remove noise or deal with low-contrast images [Di Lazzaro et al., 2013, Juang and Wu, 2011]. These pre-processing techniques make images more susceptible to feature extraction. Ultimately, feature extraction is a more pivotal part of image processing because it is the step in which semantic content is derived from raw pixel values. There are many different measures that can be used to describe an image. Features can be thought of as low level or as high level. Low-level features are descriptions of the image at the pixel level and include features derived from

statistics or from geometric transforms. High-level features are descriptors of the semantic content of an image and often require the use of heuristics.

3.1.2 Low-level Features

Some common low-level features are related to color. Color features can be extracted under a grayscale channel, different color channels or an integrated average of color channels. One way to describe a matrix of pixel values (color values) is through first order statistics (FOS). FOS are measures taken from a set of data, in our case a matrix of pixel values that do not take into account the arrangement of the data, just the collection of values. In fact, statistics are often computed from histogram representations of images (Lei et al.; Lin et al.). They include mean, standard deviation, entropy, third and fourth moments, among others (Penatti, Valle, and Torres). FOS provide a profile of the images pixel values and can applied to any image so are very universal. They are also very powerful and have been used to classify diverse image sets (Aggarwal and K Agrawal; Gonzlez-Rufino et al.).

Texture features, although they are low-level, can account for spacial arrangement. They aim to capture the changes in pixel values over the image; they measure its landscape. Many texture features try to capture the differences between adjacent or spatially related pixels or the co-occurrences of pixel values. Directional and gradient information is often maintained. There are some common feature sets used to capture texture, for example Harrilack features, co-occurrence matrices, or Laws texture energy measures . Laws measures are a set of masks that can be convolved with an image in order to detect various texture patterns. The masks allow FOS to be used to capture texture data by first transforming the image into a texture image. The masks can be 3, 5 or 7 pixels in length ad can be multiplied with each other to form two-dimensional kernels. There are

masks designed to detect level, edge, spot, and ripple relations between pixels (Eq. 3.1.13, Eq. 3.1.14 and Eq. 3.1.15) .

$$S_{level} = \begin{bmatrix} 1 & 6 & 15 & 20 & 15 & 6 & 1 \end{bmatrix} \tag{3.1.13}$$

$$S_{edge} = \begin{bmatrix} -1 & -4 & 5 & 0 & 5 & 4 & 1 \end{bmatrix} \tag{3.1.14}$$

$$S_{spot} = \begin{bmatrix} -1 & -2 & 1 & 4 & 1 & -2 & -1 \end{bmatrix} \tag{3.1.15}$$

Using all of Law's masks in combination with each other (e.g. $S_e x S_e, S_e x S_s$) provides a complete texture profile. Texture features are often used for image classification (Figure 3.1.2). They are more powerful than color features because they retain special information and images can differ greatly in their pixel distribution eve if the set of pixels for each image is identical [Jeyapoovan and Murugan, 2013, Hosseinpour et al., 2014, Kistner et al., 2013, Ursani et al., 2008]. Color and texture features are often used in combination as they account for complementary and not overlapping information about an images pixel value profile and pixel spacial distribution respectively [**?**, Penatti et al., 2012]. These low level features are universal and have been used to classify images in many fields including medicine, microscopy and food production [**?**, Jeyapoovan and Murugan, 2013, Hosseinpour et al., 2014].

Edges are another low level feature. They are yet more specific than texture. Edges are simple features of an image, but they can indicate objects and highlight the differences in image composition and are used as to classify images [Stehling et al., 2002]. An extension of edge detection is line detection. Because the kernels used in to detect edges are local, they do not look for continuous lines but rather points of local contrast. The Hough Line Transform can detect straight lines in an edge image [Szeliski, 2010]. The quantity, angles

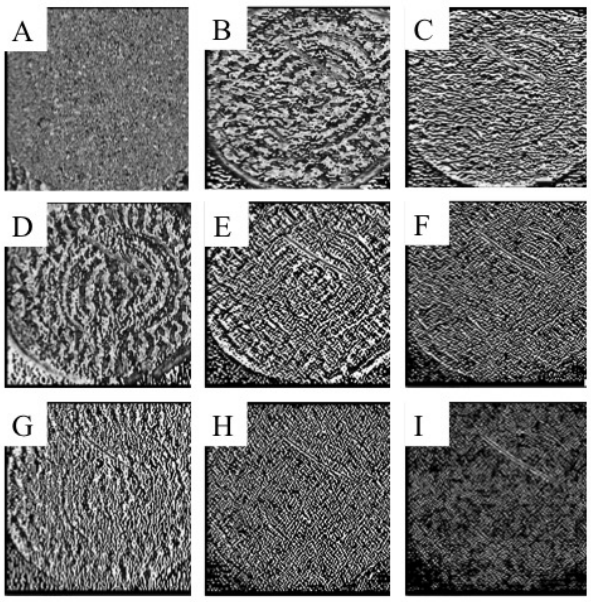

Figure 3.1.2. Law's texture energy level masks in two-dimensions. (A) level x level (B) level x edge, (C) level x spot, (D) edge x level, (E) edge x edge, (F) edge x spot, (G) spot x level, (H) spot x edge and (I) spot x spot.

and locations of lines can be informative. The Hough Line transform goes through every possible line parametrized by slope and relation to the origin. If there is an edge pixel present on the line, that line receives a vote. A threshold can be set to determine the number of votes needed before a line is detected. Similarly to detecting straight lines, the Hough Circle transforms and detects circular edges within an image where circles are parametrized by a radius and coordinates for a center, (x,y). In the transform, every point on an edge corresponds to a circle in the resulting image. In the final image, circles are determined by the points where circles intersect (Figure. 3.1.3).

The Hough Circle transform detects circles at any position with any radius. The Hough Circle Transform is considered low-level feature because it does not require any heuristics. However, it tends to describe objects and shapes in images providing higher

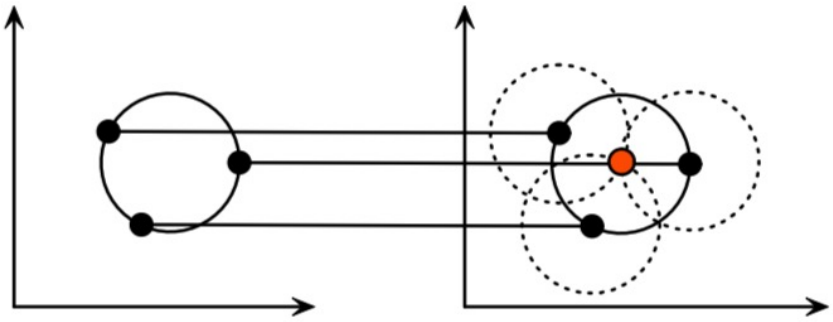

Figure 3.1.3. In the Hough Circle transform, points on an edge image get to vote for circles in the transformed image.

level content information. There are many more low-level transforms that can capture different qualities of a picture but extend beyond this introduction [Orlov et al., 2008].

3.1.3 High-Level Features

High-level features are measures of semantic content in an image. They can be more descriptive because they are tailored to the data set at hand, but they are also less universal and it is usually harder to extract. One method to extract high level features is template matching. In template matching, a mask contains an image or segment of an image that one hopes to find in the image being analyzed. This template can be convolved with an image. At each step, the difference is computed between the pixels of the image and the pixels of the template. The regions that have a high similarity to the template constitute detection [Szeliski, 2010]. The feature extracted form a detection even could be presence

or absence, or as a feature of the location of the item of interest. Different sized templates can be used to detect objects of different scales.

Another type of high-level feature is the measure of connected components. Connected components can divide an image up into its different parts. These parts can then be counted or compared to each other, their locations or sizes can be marked as features, or they can be used as masks of a region of interest. To get a good feature out of connected components analysis, one would likely have to know what type of components to expect and how to expect those components to vary between images. One study used components of an eye to detect retinal blastoma. The researchers divided the eye into connected parts and then measures the components relative sizes and positions to find abnormalities that helped diagnose cancer Haleem:2013ej.

A Fourier transform can be used to describe an oscillating pattern in an image. Fourier analysis can determine the frequency of the signal in the image, which is an indication of the variation within the image. A two-dimensional Fourier transform an also determine the periods of signals projecting in different orientations throughout an image (Figure 3.1.4). This trait would be clearly seen and only substantially useful, in images with horizontal or vertical striation.

The Hough Circle transform can detect smooth, round circles but, in images, round objects are often not perfectly round but rather they are generally round in shape. A high-level feature could explore the intricacies of contours that are irregular. For example, the convexity of a contour provides a measure of its smoothness. Convexity can be measured when a contour can provide a convex hull in an image and then the defects that are in that contour can be quantified. One way to do this for a circular contour is by measuring the variation in radii length for different angles around the center of the circle (Figure 3.1.5). This feature can be extracted using a lot of the same techniques used to gather low-level

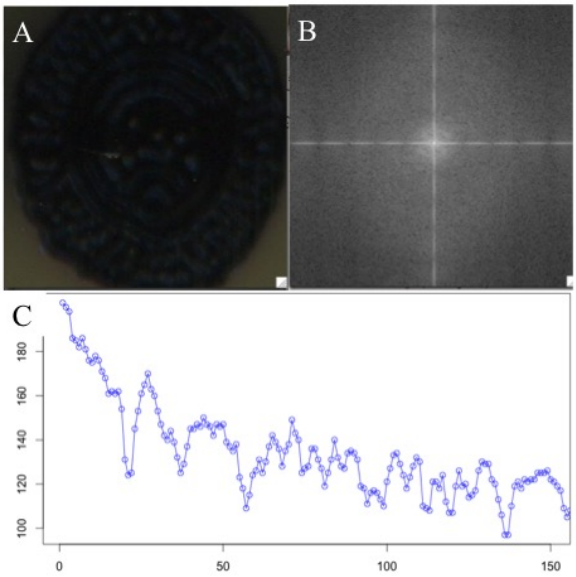

Figure 3.1.4. A Fourier transform can determine the frequency of a signal in an image, which can be a useful high-level feature. (A) original image, (B) two-dimensional Fourier transform, (D) one-dimensional slice where peaks indicate the period of the original signal.

features but convexity constitutes a high-level feature because it is only useful if images are known to have contours and potential convex-hull elements.

3.1.4 Image Segmentation

Image segmentation is often used to extract content away from a background before features are collected. This ensures that any feature data is only of the subject and not background artifacts. Image segmentation separates the foreground from the background, focusing the analysis on the content-based objects. Often heuristics provide an idea of what these foreground shapes will be, but low-level features can also be used for background-foreground segmentation.

Color can be used to distinguish a background from a foreground. Grouping methods such as K-means analysis or mean-shift methods can find regions of similar colors and

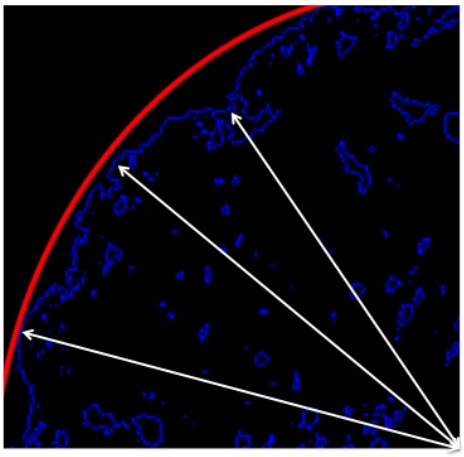

Figure 3.1.5. The convexity of a contour (value line) can be measured by the variance in the length of radii (white arrows).

determine distinctions between objects and background [Huang and Zeng, 2015, Oros-Peusquens et al., 2011,Zheng et al., 2014]. Texture can also be used to separate an image. If a texture mask is applied to an image, the result can be subjected to the same methods used upon color-based segmentation [Lee et al., 2006, Ursani et al., 2008]. Image segmentation often uses features of an entire image in order to find a submatrix within that image so that features can be re-extracted from a more salient region of the image.

3.2 Clustering

Image processing can result in large sets of data. Images can be relatively large matrices and many features can be extracted from a single image. Further, sets of images can be large. After image data has been transformed into feature vectors, it can be a

challenge to analyze the data set. Clustering techniques can be used to automatically analyze large sets of data.

Clustering is used in many branches of science and social sciences. Some of the earliest clustering algorithms emerged for analyzing sociological data. These data sets included mostly categorical data. Clustering has grown to be used with numerical and heterogeneous data sets as well.

There are different methods of clustering, but all share the basic goal of separating a data set in order to optimize the ratio of intra-group variation to inter-group variation. In other words, to find a pattern inherent in a data set. On one extreme, if every element in a data set were placed in its own group, it would have perfectly minimal intra-group variation and the pattern would be very fine grained. On the other extreme, if all the data points were one of two possible values, they would have the maximal inter-group variation and the pattern would be very corse. Clustering methods aim to find an optimal balance between the two.

Data sets can be looked at from different perspectives. Hierarchical clustering is a top-down perspective that combines the data in tiers, like a tree. Partition based clustering is a bottom up approach that initially divides the data into clusters and then adjusts the membership of those clusters.

3.2.1 Hierarchical Clustering

Hierarchical clustering successively combines data into decreasingly similar groups and thus assumes there is a branched, tree-like structure to the data set. It determines clusters using pair-wise difference measures between all the elements in a data set. Upon initialization, each element belongs to its own cluster. Clusters are then successively combined until only a single cluster remains [Murtagh, 1983](Algorithm 3.2.1).

Algorithm 3.2.1. Hierarchical clustering

1: **for** each element in data **do**
2: assign to its own cluster
3: **end for**
4: **repeat**
5: compute pairwise dissimilarities
6: combine the two least dissimilar clusters
7: **until** all elements are in a single cluster

◊

There are different ways for determining which clusters to combine, for example partial, single or average linkage update formula. Wards method uses an update formula,σ, based on variation within the possible combination of clusters, $c1$ and $c2$ (Eq. 3.2.3) between clusters [Murtagh, 1983] (Eq. 3.2.1). It chooses the cluster pair to merge that minimizes the cumulative distance between each element in a cluster. The update formula can be based on heuristics of the problem at hand, if they are known, but Ward's is a standard that can be generically applied.

$$\delta(c_1, c_2) = \frac{|c_1||c_2|}{|c_1| + |c_2|}||c_1 - c_2||^2 \tag{3.2.1}$$

An advantage of hierarchical clustering is that the final number of clusters does not need to be known a priori. However, after the algorithm is finished, it has to be decided which clades of the tree will be considered clusters. This can be done with dissimilarity measures and statistical tests, like the Students t-test, that determine when clades are significantly different from their ancestors .

This method is very costly because of determining the pair-wise dissimilarity matrix, which is $O(N^2)$. However, implementations that use reciprocal nearest-neighbor algorithms based on graph and sub-graph representations of the data set speed up the algorithm [Murtagh, 1983]. Hierarchical clustering is especially useful on smaller sets of data that have a more intricate architecture that involves different levels of similarity.

3.2.2 Partition-based Clustering

One of the most widely used partition-based clustering algorithms is the k-means algorithm. The k-means clustering algorithm was introduced by MacQueen in 1967 as an economical solution for problems involving the classification of multi-variant observations (MacQueen 1967). It divides a data set into k clusters where k must be specified by the user. Its relative simplicity likely contributed to its popularity and, today, it is a standard algorithm used in image processing. A common version of k-means is an iterative k-means such that the algorithms repeats until convergence and no elements are assigned to a new centroid in the last iteration (Algorithm 3.2.2).

Algorithm 3.2.2. K-means

1: **for** k centroids **do**
2: initialize position
3: **end for**
4: **repeat**
5: **for** each element in data **do**
6: assign it to the nearest centroid
7: **end for**
8: **for** k centroids **do**
9: recalculate coordinates as means of element's

10: **end for**

11: **until** convergence

◊

The simplest version of k-means initializes its centroids at k random points within the feature space of the problem. The results of a k-means clustering are influenced by the initial centroid locations because the solution can converge at local maxima [Bubeck et al., 2013]. To overcome this bias, one could repeat a k-means partition, choosing random centroids each time. This would prevent random error from determining the results, but could be costly in computing time. Some methods aim to distribute initial centroids in dense areas of the data [Celebi et al., 2013] . Others place centroids in different Gaussian distributions within the data. One finds centroids by finding the axis with the most variance and initializing centroids at successive medians [Alrabea et al., 2013]. Further, k-means can be combined with other algorithms like, hierarchical clustering, in order to overcome the sensitivity to local optima [Liao et al., 2013] . Methods to initialize centroids are still being explored [Celebi et al., 2013,El Agha and M Ashour, 2012]. There are lots of different methods developed to initialize centroids, but none as ubiquitous as the original k-means algorithm itself.

When using the k-means method it is important to consider its limitations. The algorithm itself is $O(N^k)$ because at each iteration the distance between each element and centroid must be taken. Time can also be increased when many values of k are experimented with. Note that k is fixed and so the algorithm is essentially linear. Further, although not significant in Big O terms, there is a coefficient of t where t is the number of iterations. Practically, a limit can be places on the number of iterations and k-means can be efficiently applied to large sets of data. The restrictions on k and t make k-means an efficient clustering option.

K-means has been used widely in image processing. It is used to cluster data represented by feature vectors. Images also inherently have patterns in them otherwise they are just random pixel values. Since clustering methods attempt to find inherent patterns, and images most likely contain patterns, they are a natural pair. It can be used for standard challenged like image segmentation and movement detection [**?**]. In these scenarios k-means is applied to cluster sections of an image. K-means has also been used to cluster images based on higher-level feature values and is used in image retrieval and recognition [Stehling et al., 2002] . As long as images can be represented by feature vectors either direct pixel values or extracted features then they can be subjected to k-means clustering.

One strength of k-means is its generality and how widely applicable it is. It is used in fields as diverse as, social science, health science, and computer vision. In complicated real-world problems, it is often a solution to a sub-problem and used with other tools. It is common because there is often no training data available and so it is used to explore, maybe early stages of a study.

Clustering is often supported by other methods of analysis. For example, hierarchical k-means clustering algorithms are designed to harness the strengths of both methods [Liao et al., 2013]. Two clustering methods can be used in tandem in order to boost our confidence in the results. Or k-means can be used with other classification techniques, such as supervised machine learning [Zheng et al., 2014] . These ensemble methods can be bundled into new versions of k-means (e.g. warped k-means, k-means with gravitation, iterative Fischer, and constrained k-means) [Clausi, 2002,Hu et al., 2008,Leiva and Vidal, 2013]. The breadth of different k-means variants is beyond the scope of this review, but the fact that that breadth exists speaks to the utility of k-means as a tool

3.2.3 Feature Weighting

An inherent limitation to clustering solutions is the reality of the data; some data sets will be more homogenous than others. Of course, in real-world problems, this measure of usefulness is also a measure of the usefulness of the feature variables used to cluster and thus the efficacy of the clustering alone is not obvious. Clustering seeks to find architectures in data, but it cannot create them if the do not exist. Because there is often noise in real-world data sets, clustering algorithms can be augmented to be more robust.

Feature weighting can be used to encourage patterns to emerge from data. Modifications of clustering algorithms have been written to factor different features with different weights [Modha, 2003, Cordeiro de Amorim and Mirkin, 2012]. In particular, one method gives increased weight to features that show the most variance between clusters [Huang et al., 2005]. Weighting such a feature more heavily makes it more likely that a cluster will form around the pattern exhibited by that single feature. Since that feature is chosen because it varies in different clusters well, the final clusters will be more likely to have that strong variance reflected in its final variance. This method could be useful to prevent relatively uniformly distributed features, or noise, from clouding cluster architectures.

3.2.4 Assessing Clustering Solutions

In order to assess a clustering solution, one can look at intra and inter class variation. Sum of squared error (SSE) is a difference measure for all clusters, C_i of clustering solution based off of Euclidean distance that can be used to calculate variation between two points, i and j in p dimensional space (Eq. 3.2.2 and Eq. 3.2.3)The percent of the total variation explained by the inter-class variation provides a numerical measure of the solutions strength. Some measures of a successful clustering solution are a low intra to interclass variation ratio or a high inter to total variation ratio. However, even examining

these numbers has a limit to its usefulness because there is no standard of good and bad clusters; the optimal solution depends on the data being worked on.

$$dist^2(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + ... + (x_{i_p} - x_{j_p})^2} \tag{3.2.2}$$

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_1, x) \tag{3.2.3}$$

When used to solve real-world problems, clustering solutions can also be assessed by their usefulness: whether the partitions created have some practical, functional value specific to the task at hand. This is not so easy if the data set is completely foreign, but often in real-world problems, heuristics exists that help one know whether the clusters are useful.

3.3 Support Vector Machines

In the late 1970s, a Russian mathematician, named Vapnick, worked to solve the generalized pattern recognition problem. Many scientists were working on this problem because of its utility and challenge; the aim was to successfully categorize data without predefined criteria for classification. Using statistical learning theory, Vapnik put forth a method that would automatically learn the criteria of classification from sets of labeled data.

The Support Vector Machine (SVM) is a supervised machine learning tool that uses statistical information to solve classification problems. As a supervised technique, the SVM requires a training set of data complete with data points and their labels. It can then generalize the differences between elements of different classes and assign labels to novel, unlabeled data. The SVM is powerful because it can learn a trend from a small sample of data and generalize that trend to never before seen data. The SVM finds the

hyper plane in feature space that best separates the classes of the training data. It is easy to visualize this process for the binary classification of two-dimensional data, assuming data points have two features (x and y) (Figure 3.3.1). The SVM finds the line that separates the data points in the two classes from each other and has the widest margins, in Euclidian distance) from the individuals closest to the hyperplane, which are called support vectors.

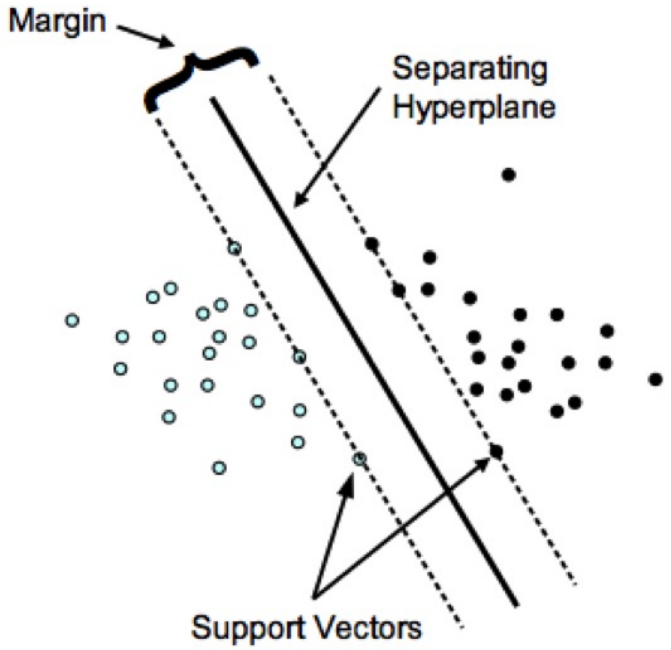

Figure 3.3.1. A Support Vector Machine finds the hyperplane that bests divides a set of pre-labeled data using only the points closest to the hyperplane. [Ben-Hur and Weston, 2009]

The ingenuity of the SVM lies in its use of support vectors, the points closest to the hyper plane, to determine the optimal separating hyper plane. Maximizing the distances from these margins thus maximizes the distances between the groups at large. Soft margins are implemented in SVMs so that outlying data can be weighted lightly and effectively ignored (and thus mis-classified) for the betterment of the classifier at large. C is a parameter that controls this. Because of this, the SVM is immune to outlier data that can cause over-fitting and make a classifier too specific and not applicable to unseen data.

The SVM is a good tool for classifying linearly separable data because good hyper planes exist between the classes. However, the SVM can be applied to non-linear data as well. Kernels have been developed to transpose data separated by a higher order functions onto a linear hyper plane. In addition to the linear kernel there are polynomial, Gaussian radial bias and spline kernels (Figure 3.3.2). Often it is difficult to know which kernel best fits a data set so different kernels need to be experimented with and compared to each other.

kernel	formula	parameters
linear	$\mathbf{u}^\top \mathbf{v}$	(none)
polynomial	$\gamma(\mathbf{u}^\top \mathbf{v} + c_0)^d$	γ, d, c_0
radial basis fct.	$\exp\{-\gamma\lvert\mathbf{u} - \mathbf{v}\rvert^2\}$	γ
sigmoid	$\tanh\{\gamma\mathbf{u}^\top \mathbf{v} + c_0\}$	γ, c_0

Figure 3.3.2. Kernels can transpose data that is not linearly separable into linear space so that it can be analyzed by an SVM. [Ben-Hur and Weston, 2009]

In addition to classification, SVMs can be expanded to multiway classification. Multi-class classification problems can be redesigned as a series of binary classification problems. The ordering and particulars of these binary classifications may depend on the specific classification problem. A common technique is a series of 1 vs. ALL tests that

train on each group vs. the other groups combined . Integrated algorithms for one-step multi-class SVMs have also been developed [Gidudu Anthony and Tshilidzi,].

The statistical theory that inspired SVMs has been adapted for unsupervised learning in the form of Novelty Detection or one-class SVMs [Hyduke and Palsson, 2010]. Single class SVM uses a kernel, usually radial, to determine a distribution within a feature space. The distribution can be trained on a single class of data, and elements that fall within that distribution in test data are classified as part of that class. Similarly to two-class SVM, this is sensitive to the parameters associated with the kernel shape(in the case of a radial kernel, nu controls sensitivity) and C, the penalty parameter. Unlike clustering methods, the one-class SVM allows for unsupervised learning that only classifies a subset of the data. The solution does not need to take into account the architecture of the whole data set, it can just find a pattern within it. When you have a class and elements that dont belong to that class, but you know there is variability in those elements, and they likely should separate from each other, a classic binary SVM approach would not be ideal because the distribution on one side of the hyperplane will be too variable.

The performance of an SVM can be assessed by the number of correct and incorrect classifications. These include true positives and true negatives compared to false positives and false negatives. Because SVMs are supervised and require a set of labeled data, that data can be used for assessment. Often the data is divided into a training set and a test set. It is also common to analyze performance through cross-validation accuracy wherein a small subset of the labeled data is held out for training and testing is applied to the rest repeatedly selecting different subsets of data to hold-out. Grid search uses cross validation accuracy to assess the performance of SVMs that are run on the same data with a ranges of parameters.SVMs can be very sensitive to parameters C, gamma and others. Grid search is a common algorithm to help determine the most useful parameters. There are extensive theories on how best to optimize an SVM [Shawe-Taylor and Sun, 2011].The results of

grid search are one dimensional if a linear kernel and two-dimensional taking into account the parameter of the kernel. Grid search runs SVMs with ranges of parameters and return the parameter set with the highest cross-validation accuracy.

With the ever-increasing presence of the internet and importance of computer vision and image analysis, the SVM is frequently utilized to classify images. An image that is represented as a vector of features or, equivalently, a point in multi-dimensional feature space is susceptible to SVM classification. In real world applications, there is not often a problem 'clean' enough that a traditional two-class SVM is enough to provide interesting results. Often, like k-means, SVMs are used in compound ways or in conjunction with other classification methods to build ensemble classifiers. Two SVMs were used on the same set of images, one to classify based on low level features and one to classify based on high level features and then the results of the two machines were combined in a weighted manner in order to make the final decisions on the image classes [Qi and Han, 2006]. There have even been pipelines involving four machines Saha:dc. SVMs have been adapted for multi-way classification using a series of one vs. all classifiers; the supervised SVM technique is beginning to be combined with unsupervised k-means clustering deNazareSilva:gh. The complexity inherent in many real-world computer vision and image processing tasks requires the creative use of the SVM tool.

The power of an SVM to classify large amounts of data makes it a good tool for handling image data, including that from high-throughput functional assays, especially those that collect image data. Although images of a microbial mutant library do not come with labeled data and cannot be immediately classified by an SVM, once a phenotype of interest has been identified, an SVM can be used to quickly screen an entire library for similar mutants. This is critical, especially when working with saturation libraries, in order to collect all possible genes that contribute to a complex trait, such as biofilm. Tools like clustering and SVMs allow for the exploration of an organism from the point of view of

an interesting phenotype instead of a single gene. This allows for more useful and efficient exploration of genomes such as those from environmental isolates.

4
Materials and Methods

4.1 Medias and Strains

Environmental isolates and transposon mutants were maintained on TGHL agar (0.2% lactose, 0.4% gelatin, 1.6% tryptone, and 1% agar) plates at 22^oC and stored at -80^oC in 25% v/v glycerol and R2A (0.31% VWE R2A mix) stocks. During selection for transposon insertion, TGHL was supplemented with kanamycin in 45 45 mg/mL. *E. coli* strain $DH5\alpha\lambda pir$ was maintained on LB (0.5% NaCl, 0.5% yeast, 1% tryptone and 1.5% agar). supplemented with Km45 and incubated at 37^oC. Swimming and swarming motility assays were amended with 0.3% and 0.5% agar, respectively.

4.2 Genomics

Next generation Illumina technology was used to sequence whole genomes. Specifically, sequencing was done on a MiSeq with 100 bp paired-end reads. Resulting reads were assembled into contigs using MIRA (Mimicking Intelligent Read Assembly), annotated with RAST (Rapid Annotation Using Subsystems Technology) and visualized and

subjected to comparative analysis in the SEED viewer. BLAST (Basic Local Alignment Search Technology) search was used to further explore genomics.

4.3 Mutagenesis

E coli strain WM1590 was used as a donor strain in conjugation. Mating strains were struck out onto LB agar plates and incubated at 22^oC for 48 hours. Resulting colonies were then suspended in R2A broth and diluted to a 2.5 McFarland, $100\mu l$ of culture were plated onto R2A/Km45 plates and the rest of cells were centrifuged for 3 minutes at 13,000 rpm, re-suspended in $200\mu l$ R2A and spread with beads onto R2A/Km45 plates.

4.4 Biofilm Morphology Assay

From selection plates, 24 hour old colonies were transferred to $100\mu l$ R2A/Km45 broth, incubated at 22^oC , shaking for 48 hours then plated in $2\mu l$ samples onto square TGHL plates in a 6 by 8 pattern. Liquid cultures of each environmental isolate and mutant were brought to 25% glycerol and frozen at -80^oC . Plates with lids removed were scanned after 3,4 and 5 days on an Epson300 flatbed scanner with a white background at 1200 dpi in jpeg format.

4.5 Image Segmentation

Scans of the mutant libraries were processed with scripts written in C++ using the OpenCV library. The original color images were converted to grayscale, eroded, dilated, blurred and then converted too binary images using the Otsu method. The Hough Circle Transform was used to detect colonies within the image (minimum distance: 400, edge detection threshold: 250, center detection threshold: 25, minimum radius: 100 and maximum: 250). The boundaries of sub-matrices were determined by the detected Hough circles with

a padding of 20 pixels in each direction, except for edge cases, which were cut off at the boundary of the image. Sub-matrices were saved as individual images.

4.6 Feature Collection

Five sets of features were collected: colony features, color features, texture features, convexity features and Fourier transform features (Figure 4.6.1).

Class	Feature	Definition
Colony Features	MinValue	$\min(\{i \mid P(i) > 0\})$
	MaxValue	$\max(\{i \mid P(i) > 0\})$
	ColonyHeight	$\lvert G_y \rvert - 40$
	ColonyWidth	$\lvert G_x \rvert - 40$
	PixelCount	$\lvert \{i \mid i > 0\} \rvert$
Color Features	Mean	$\mu^c = \sum_i iP^c(i)$
	StandardDeviation	$\sqrt{\sum_i (i-\mu^c)^2 P^c(i)}$
	ThirdMoment	$m_3{}^c = \sum_i (i-\mu^c)^3 P^c(i)$
	FourthMoment	$m_4{}^c = \sum_i (i-\mu^c)^4 P^c(i)$
	Entropy	$H^c = \sum_i P^c(i)\log(P^c(i))$
Texture Features	Mean	$\mu^{G\otimes t} = \sum_i iP^{G\otimes t}(i)$
	StandardDeviation	$\sqrt{\sum_i (i-\mu^{G\otimes t})^2 P^{G\otimes t}(i)}$
Convexity Features	MeanDistance	$\mu^r = \sum_{r\in R} rP(r)$
	StandardDeviationDistance	$\sqrt{\sum_{r\in R} (r-\mu^r)^2 P(r)}$
	MinDistance	$\min(R)$
	MaxDistance	$\max(R)$
	RangeDistance	$\max(R) - \min(R)$
	MeanEdge	$\mu^e = \sum_{e\in E} eP(e)$
	StandardDeviationEdge	$\sqrt{\sum_{e\in E} (e-\mu^e)^2 P(e)}$
	MinEdge	$\min(R)$
	MaxEdge	$\max(R)$
Fourier Features	Dip 1	1st local min in 1D slice of G
	Peak	1st local max in 1D slice of G
	Dip 2	2nd local min in 1D slice of G

Figure 4.6.1. Feature definitions.

Where, G is histogram of image C is the channels of G P is the probability histogram T is the set of texture masks L = level mask = E = edge mask = S = spot mask R is the set of radii around a colony E is the set of edges used to measure radii, i G,

$i \in G$, $C = \{red, green, blue\}$, $c \in C$, $P(i) = \frac{G[i]}{\sum_i G[i]}$,

$T = \{LxL, ExE, SxS, SxE, LxE, LxS, ExS, ExL, SxL\}$, and $i \in T$.

Additionally, the path to the described image was included as a feature in each set of features (but not used for clustering or classification). The set of features was a comma separated value (csv) sheet generated by iostream output from C++ scripts. The feature set was loaded into R as a data frame. This data frame was normalized to values between 0 and 1 according to the function where and D is the data frame, and Y is the set of FileName features (Eq. 4.6.1).

$$f(x) = \frac{x - min(D)}{max(D) - min(D)} \tag{4.6.1}$$

4.7 Feature Selection

All sets of features, (colony features, color features, texture features, convexity features and Fourier transform features) were used in all combinations in the k-means method from the cluster library in R. Further, weighted k-means clustering solutions were also explored using R libraries. In all cases, k was chosen manually by computing results for k values from 2 to 25 and then selecting smallest k that still provided a steep reduction, noticeable by eye, in SSE from k-1.

4.8 Clustering and Classification

Clustering and classification based on the csv data sets were performed in R. Workflow went from a library of mutants to a group of mutants that share a phenotype of

Table 4.8.1. Libraries used in R

Library	Use
cluster	k-means and hierarchical cluster analysis
e1071	SVM, one-class and two-class
graphics	displaying plots
jpeg	displaying images
parallel	parallel processing used with hierarchical clustering
pvclust	visualizing clusters, PCA
stats	assessing cluster solutions
WeightedClustering	comparing clustering algorithms
weightedKmeans	weighted-feature clustering

interest in a semiautomated way (Figure 4.8.1). Libraries were used when appropriate for data normalization, k-means clustering, hierarchical clustering, SVM learning and data visualization (Table 4.8.1). The k-means algorithm resulted in clusters that were manually analyzed. Ransom samples were analyzed for colonies that looked similar, different, or obscured by artifacts. Each cluster in each solution was scored based on the maximum number of colonies that could be considered in a group, the number of colonies with plating artifacts, the number of colonies with typing over their images and the number of images that did not contain an colony. Clusters whose random sample of 24 elements was mostly images with numbers, missed colonies or plating artifacts were removed from the working data. The full data set was cleaned of mutants that resemble WT using one-class SVM novelty detection (radial kernel, gamma = 0.01, cost = 10, nu = 0.0001). The data was clustered by the k-means method and resulting clusters of interest were subsequently subjected to Wards hierarchical method (and bootstrap difference measures) when appropriate. A final resulting cluster was chosen as a focus group and an SVM was trained in two-way classification between the focus group and WT. This SVM was used to predict the classification of the original, entire data set. Grid search provided the parameters for SVM and a linear kernel was used (linear kernel, gamma = 0.03125, cost = 1).

Table 4.9.1. Measures of Biological Relevance

Measure	Explanation
in	The number of images of colonies that share the dominant phenotype
out	The number of images of colonies that do not have the dominant phenotype
number	The number of images with big, white numbers printed on them
missed	The number of images that do not contain a colony
artifact	The number of images with a disruptive mark
other	number, missed or artifact

4.9 Assessment

Clusters were assessed by intra and inter cluster variance, with intra class variance normalized by the c luster size. Random subsamples of 24 elements from each cluster were analyzed manually (Table 4.9.1). Clusters were scored using quality statistics: Point Biserial Correlation (PBC), Huberts Gamma(HG), Average Silhouette width (ASW), R2 (R2) , and Huberts Coefficient (HC) as described in the WeightedClustering in R manual. Values were computed using the WeightClustering library in R. SVM results were analyzed by 10-fold cross validation accuracy and tested on a manually labeled data set of 100 elements.

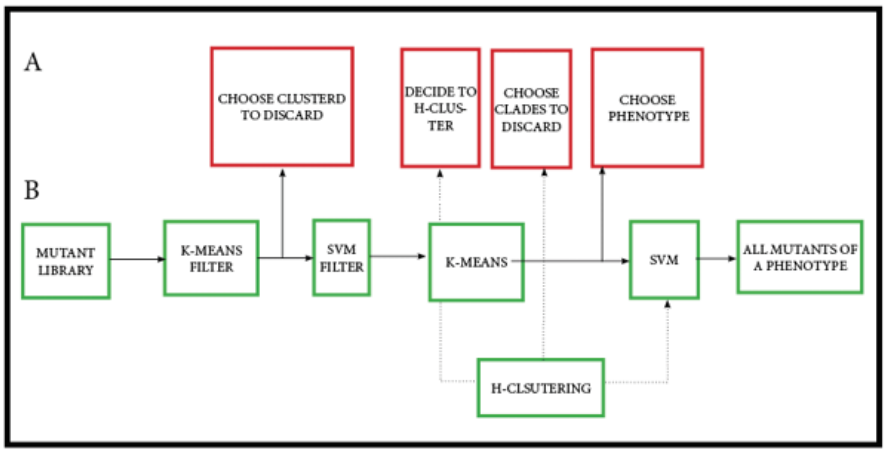

Figure 4.8.1. Workflow of machine learning methods that find a group of mutants that share a phenotype from a large library. (A) Steps that require human input to direct the machine learning methods toward a biologically relevant solution. (B) Automated steps of filtering, clustering and classification.

5
Results

5.1 BJB312 Genome Assembly and Annotation

The BJB312 genome shows 403 different subsystems. Genes suggest motility, antibiotic and metal resistance, antibiotic production, polysaccharide production and excretion, and cAMP signaling. The three most prevalent subsystems are carbohydrates, amino acids and derivates and protein metabolism. Stress response, motility and membrane transport were also clearly present.

5.1.1 Comparative Genomics

BJB312 showed varying similarity to genomes of BJB1 or BJB302, other violacein-producing environmental isolates from the Hudson Valley Watershed (Figure 5.1.1). The genome can also be looked at in comparison to genomes from available databases. The gene of BJB312 more closely resembles that of *J. lividum* relative to *J. marseille* and *E. coli* (Figure 5.1.2). It has been classified as a strain of *Janthinobacterium sp.* based on genome similarity. However, many of its close neighbors in various genuses including *B*urkholderia (Table 5.1.1).

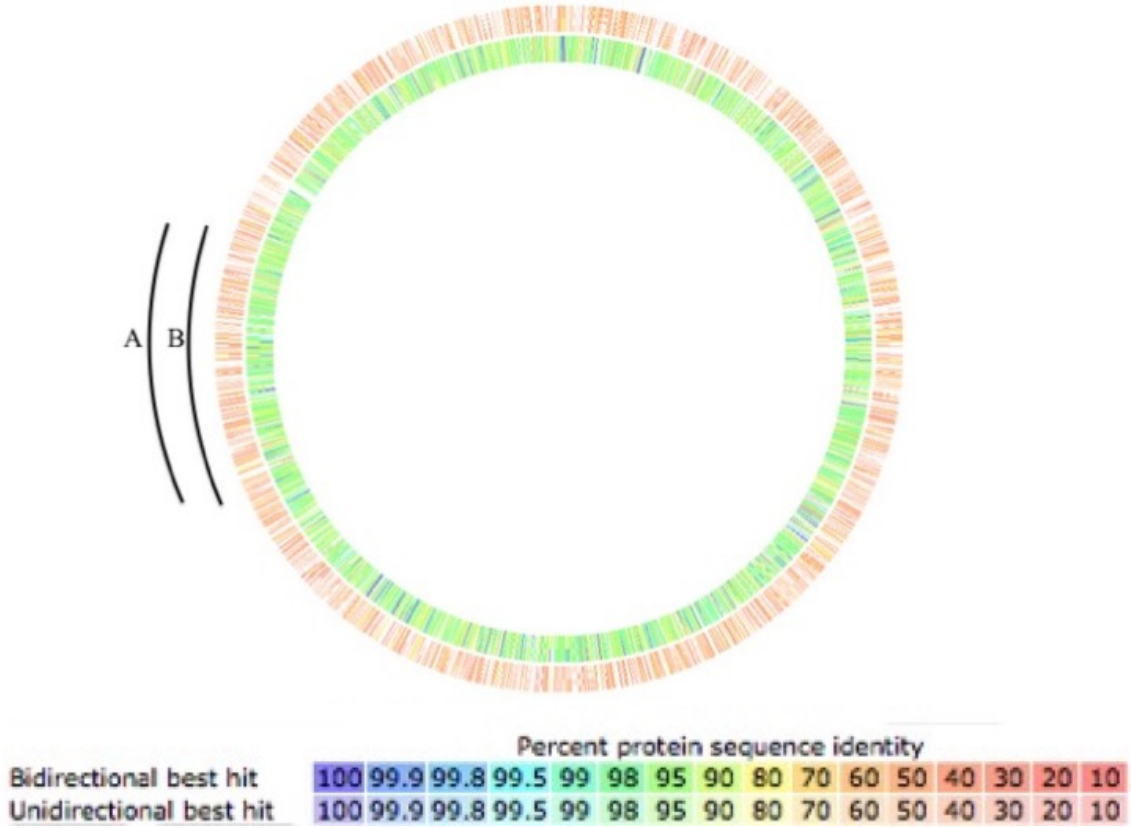

Figure 5.1.1. Sequence based comparison between BJB312 and environmental isolates (A) BJB302 (B) BJB1.

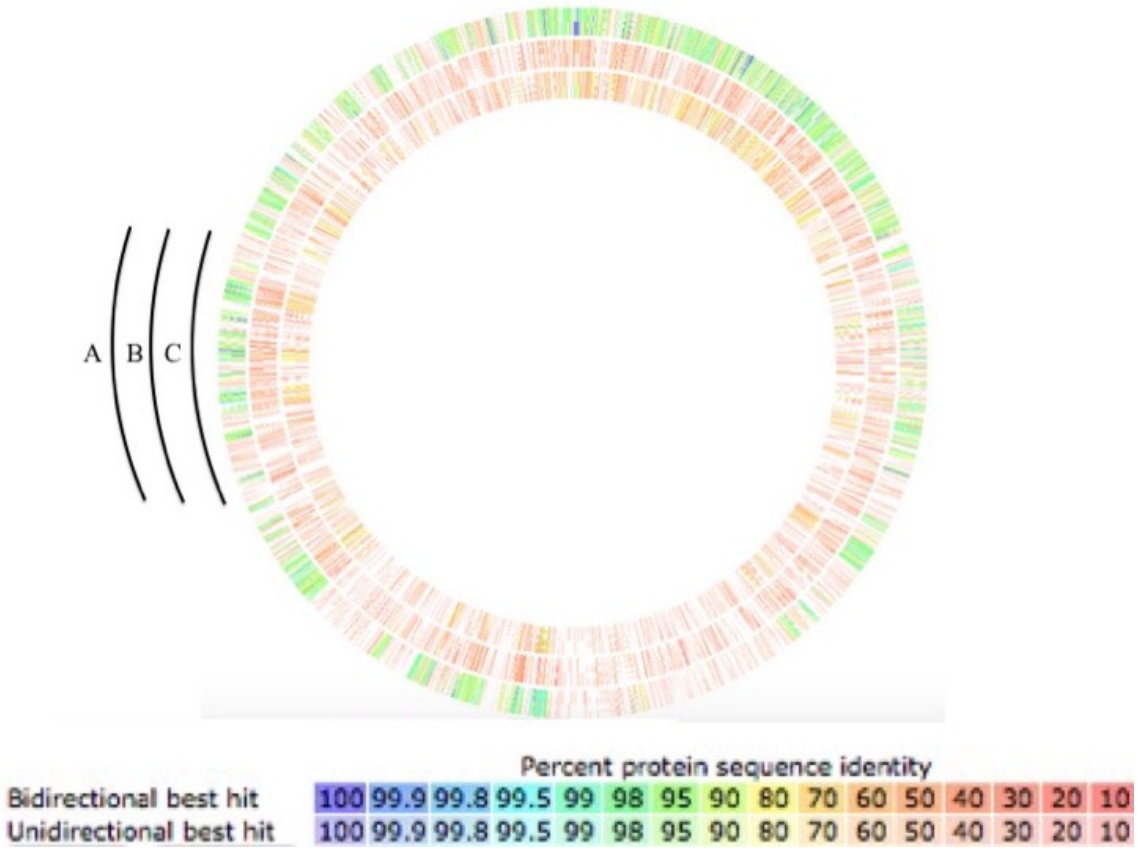

Figure 5.1.2. Sequence based comparison between BJB312 and (A) Janthinobacterium lividum (B) *Escherichia coli* and (C) *Janthinobacterium marseille* visualized in SEED.

Table 5.1.1. Closest neighbor strains to BJB312

Organism	Score
Janthinobacterium marseille	507
Herminiimonas arsenocoxydans	457
Oxalobacter formigenes	329
Janthinobacterium sp.	200
Burkholderia xenovorans	194
Burkholderia phytofirmans	179
Burkholderia sp.	170
Burkholderia dolosa	167
Ralstonia eutropha	164
Herbaspirillu sp.	163

5.1.2 Genes of Interest

The *v*io operon was found in the genome of BJB312 in a BLAST result (Figure 5.1.3). All five genes, *v*ioABCDE are present, spanning two contigs. This is the same operon found in all known violacein producers.

In addition to the presence of violacein genes, genes relating to biofilm were identified (Table 5.1.2). These genes have known roles in EPS production, flagellar motility and Type II secretion systems and suggest a system of biofilm production and regulation.

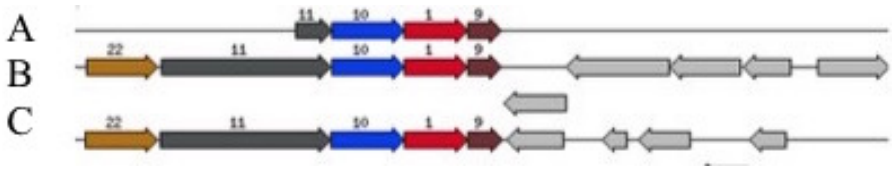

Figure 5.1.3. A section of the vio operon was found with a BLAST search in (A) BJB312, (B) *Chromobacterium violaceum*, and (C) *Pseudomonas tunicata*. For all, vioB (dark grey), vioC (light blue), vioD (red) and vioE (maroon).

Table 5.1.2. Genes of interest found in genome of BJB312

Gene(s)	Function
pel	Pectate lyase
GGDEF and EAL domains	cyclic-di-GMP
rhl	RNA helicase
luxP	Quorum Sensing
fliEGLMP,flgCBK,fleN	Flagella
flp,cpaC,repA,tadBCZ	TypeII Secretion System (pilus)

5.2 Functional Assays with BJB312

BJB312 is an environmentally isolated strain of bacteria. It is a Gram negative bacterium isolated from fresh water in the Hudson Valley region of NY state. In order to observe colony morphology and behavior, BJB312 was examined on a variety of media types. BJB312 forms biofilm pellets in static liquid cultures of LB, TGHL, 1% Tryptone and R2A broth (Figure 5.2.1). It produces biofilm when grown in shaking liquid cultures

(220 rmp) in LB, TGHL, and tryptone broth. The violacein production in R2A liquid cultures is subtle. On agar plates it produces violacein in a similar pattern across media types (Figure 5.2.2 and Figure 5.2.3). On TGHL it appears particularly rugose. BJB312 spread widely on TGHL and R2A swimming motility plates and TGHL swarming motility plates. On other swarm and swim motilities it shows only growth and moderate swarming (Figure 5.2.4 and Figure 5.2.5).

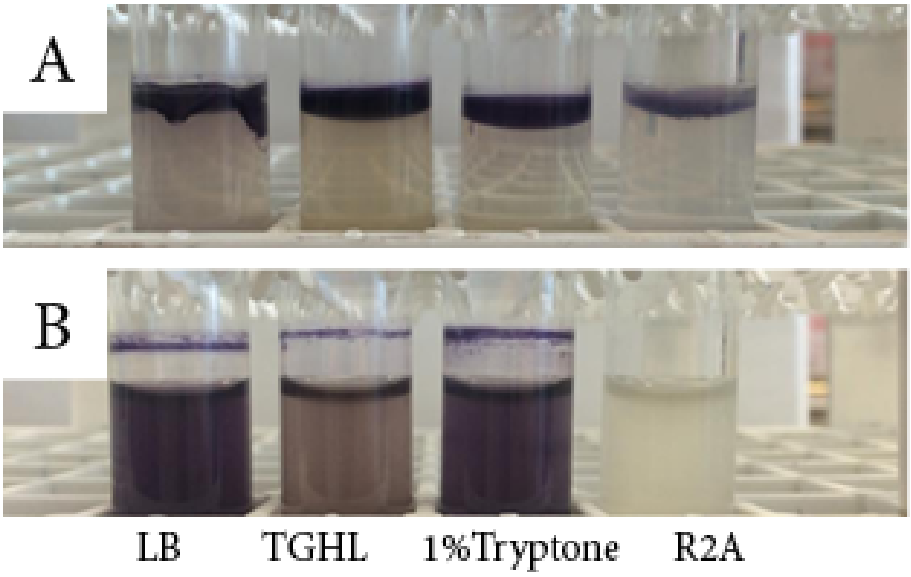

Figure 5.2.1. Environmental isolate BJB312 displays (A) differential biofilm pellet formation in static liquid culture and (B) violacein production in agitated liquid culture (B) across medias LB, TGHL, 1% Tryptone and R2A, left to right.

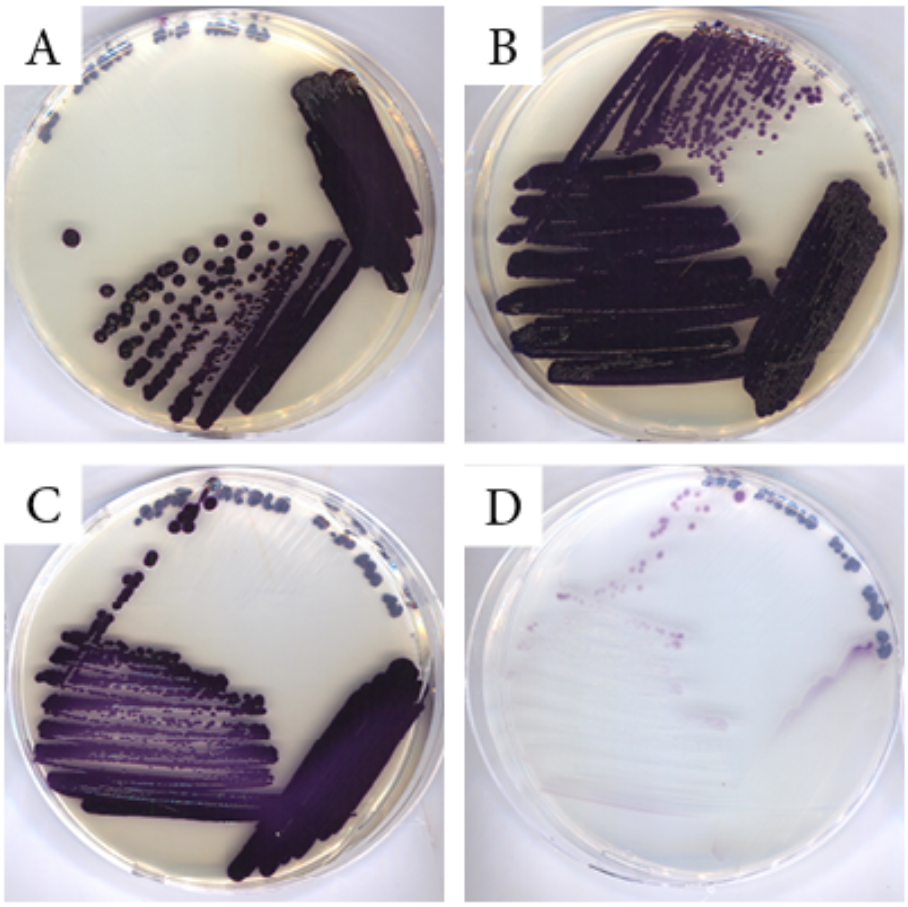

Figure 5.2.2. Environmental isolate BJB312 grown on 1.5% agar plates of the medias (A) LB, (B) TGHL,(C) 1% Tryptone and (D) R2A. Contrasted against a white background.

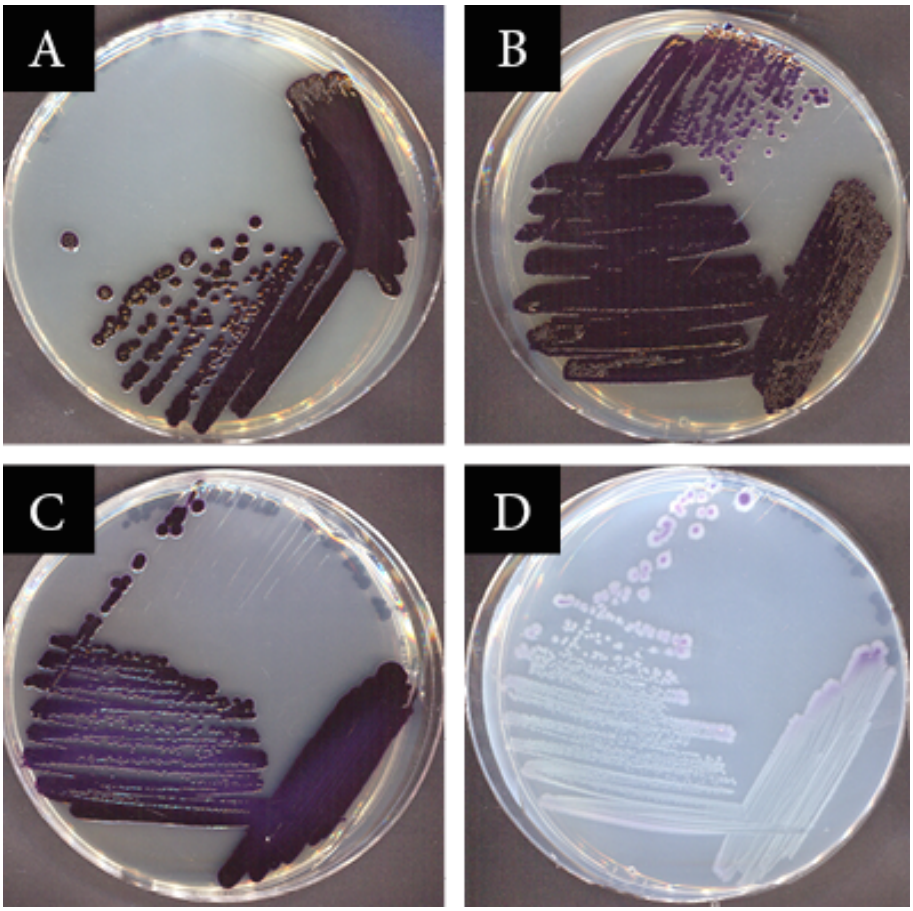

Figure 5.2.3. Environmental isolate BJB312 grown on 1.5% agar plates of the medias (A) LB, (B) TGHL,(C) 1% Tryptone and (D) R2A. Contrasted against a black background to display subtle differences.

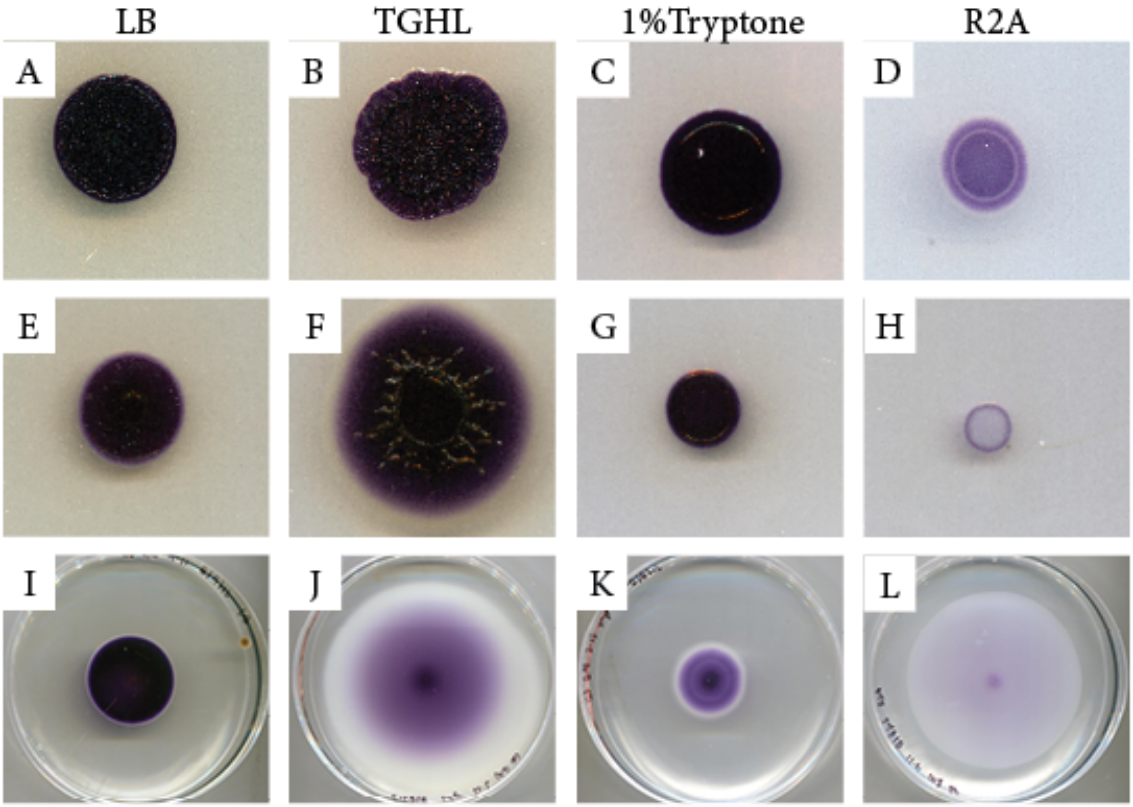

Figure 5.2.4. Environmental isolate BJB312 differentially displays stationary growth (A-D), swarming motility (E-H) and swimming motility (I-L) across four different medias LB, TGHL, 1% Tryptone and R2A, left to right. Contrasted against a white background.

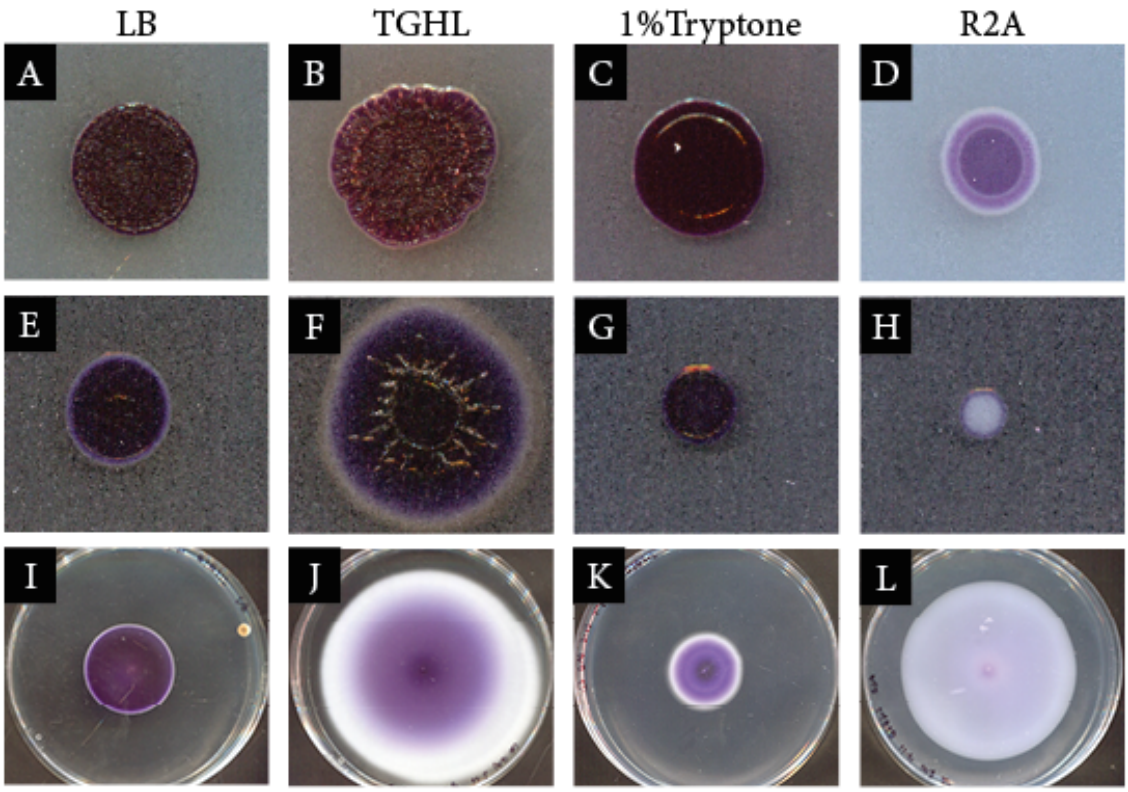

Figure 5.2.5. Environmental isolate BJB312 differentially displays stationary growth (A-D), swarming motility (E-H) and swimming motility (I-L) across four different medias LB, TGHL, 1% Tryptone and R2A, left to right. Contrasted against a black background to display subtle differences.

5.3 Image Segmentation

Images of individual bacterial colonies were successfully isolated from scans containing up to 288 colonies. The scans were susceptible to the Otsu method of binaryization. Further, the near-circular shapes of the colonies proved to be circular enough for the Hough Circle transform (Figure 5.3.1). The collected image library had a true positive value or 82.26% and a positive predictive value of 56.95%. Additionally, 238 true-positive images of BJB312 were extracted from a scan containing 288 colonies.

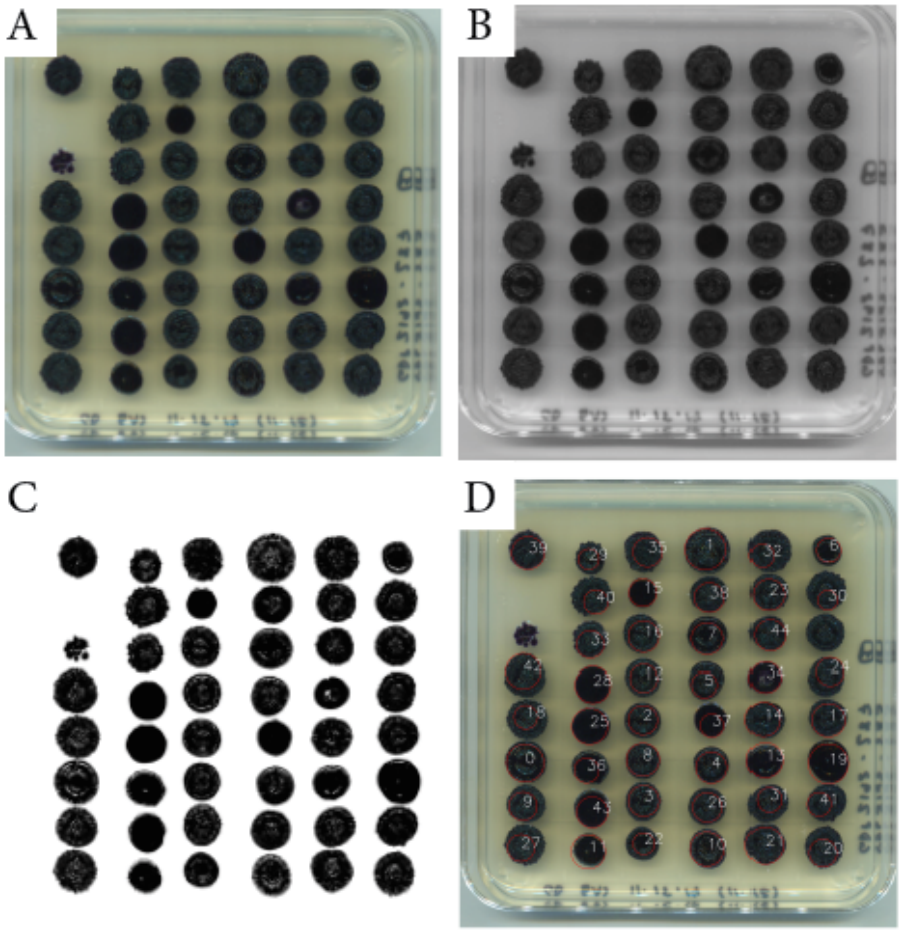

Figure 5.3.1. The original scans collected for the library had up to 6 plates in one scan with up to 48 mutants per plate. Individual mutants were isolated using the workflow: (A) original image, (B) greyscale image, (C) binary image using the Otsu method and finally (D) colonies detected using the Hough Circle transform.

5.4 Clustering Solutions

5.4.1 Machine Learning Filters

The k-means filter revealed 9 clusters that were removed from the working data sets GDB_k and GDB_{ks}. These clusters were rich with numbers and plating mistakes (Figure 5.4.2). A one-class SVM that was trained on images of WT, BJB312 colonies classified 1117 of the GDB mutants as WT, 2696 as mutant and 468 as NA. The removal of these mutants led to the determination of data set GDBs and, in part, GDBks.

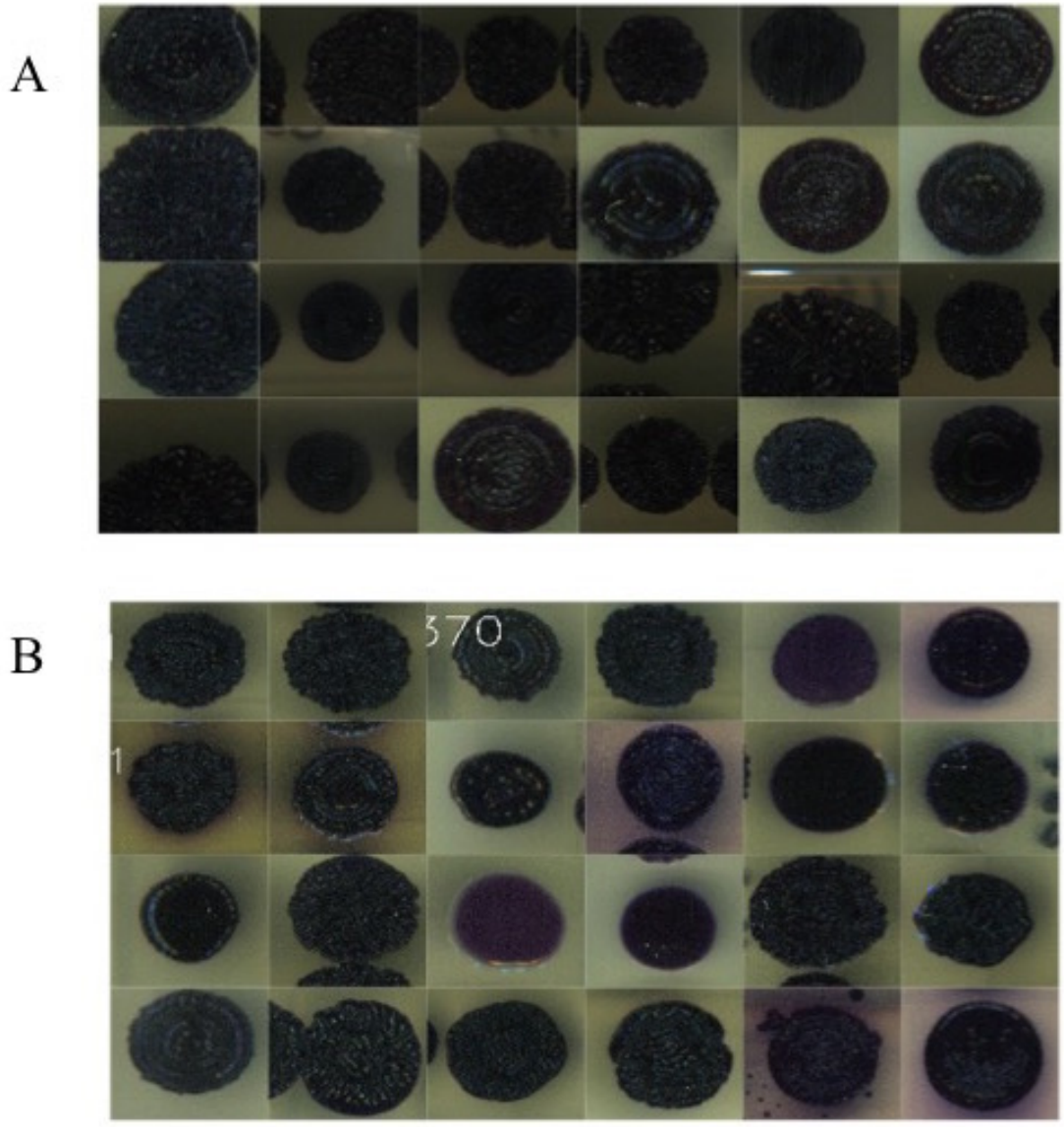

Figure 5.4.1. Samples of 24 randomly selected elements from sets of (A) BJB312, WT and (B) insertion mutants from GDB that were classified as WT using a one-class lib svm when C = 10, nu = 0.0001 and gamma = 0.01.

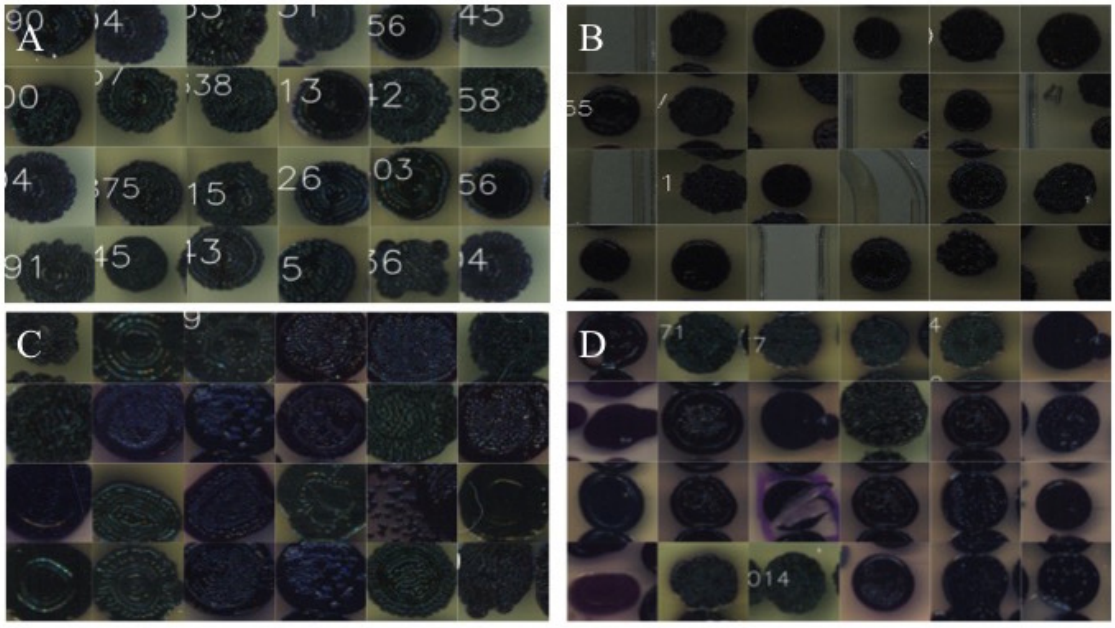

Figure 5.4.2. Samples of 24 randomly selected elements from clusters that were removed from GDB with k-means filtering. (A) In cluster 14 (size = 122) printed numbers obscured the colonies. (B) Cluster 22 (size = 48) contained many images that did not include full colonies.(C) Many of the images in cluster 4 (size = 132) only captured the centers of colonies. (D) Images from cluster 17 (size = 116) showed contamination and plating artifacts.

5.4.2 K-means Solutions PCA

When data sets of unrefined data (GDB), data refined by k-means (GDB_k), data refined by SVM (GDB_s) and data refined by both SVM and k-means (GDB_{ks}) were subjected to k-means clustering with 25 centroids, they resulted in different solutions. The differences in the architectures observed via principle component analysis differ in the distribution of clusters. The double refined data solution does not display an outlier in a way similar to the other solutions (Figure 5.4.4).

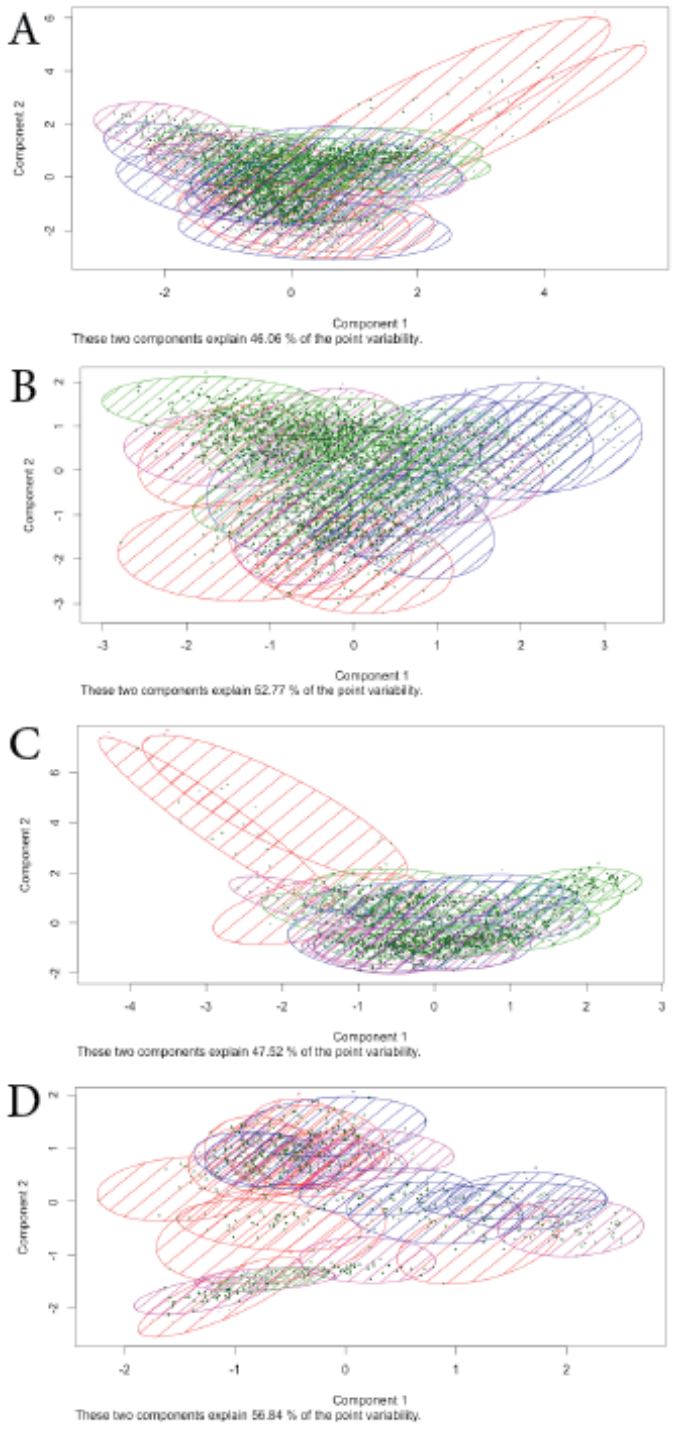

Figure 5.4.3. Principle Component Analysis visualizations for clustering solutions of k-means where k = 25 for data sets (A) GDB, (B) GDB_k, (C) GDB_s and (D) GDB_{ks}.

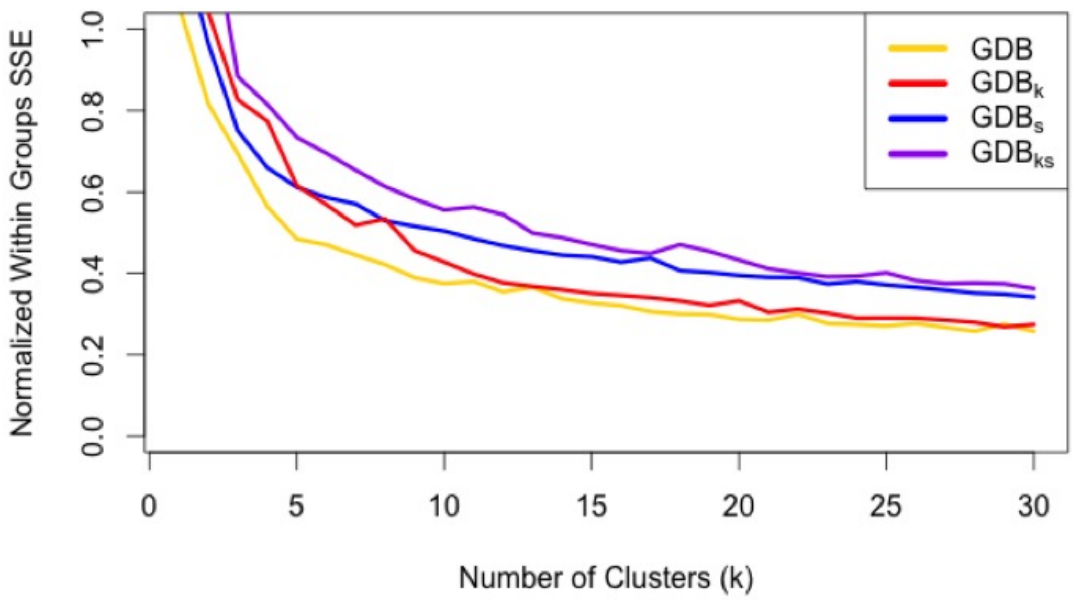

Figure 5.4.4. The total within-groups SSE divided by the size of the data set for clustering solutions with k in range (2:30) for data sets GDB, GDB_k, GDB_s and GDB_{ks}. The slope for all data sets is relatively level at k = 25.

5.4.3 SSE

The cluster solution from data singly and double refined provided a higher proportion of SSE that was between cluster SSE. This trend was observed in cluster solutions for k = 15,25,50,100 (Figure 5.4.5).

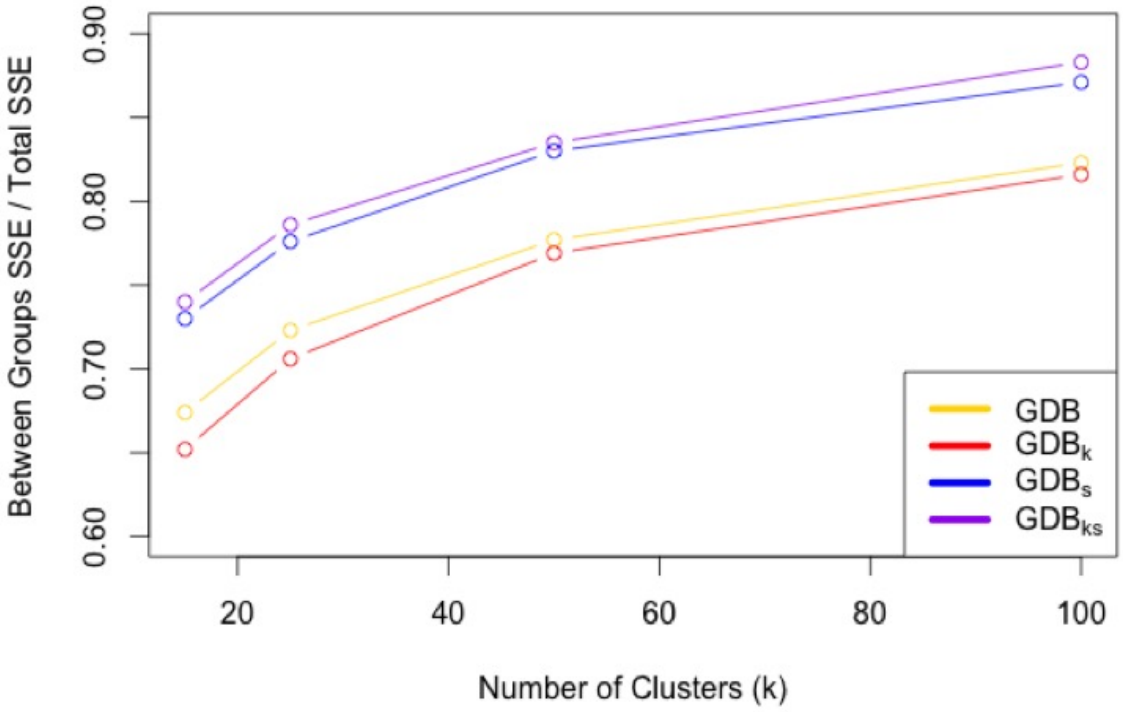

Figure 5.4.5. The ratios of SSE between groups to total SSE for clustering solutions with k = 15, 25, 50 and 100 from the data sets GDB, GDB_k, GDB_s and GDB_{ks}.

5.4.4 Quality Statistics

The clustering solutions from refined data are superior according to statistical quality measures: Point Biserial Correlation (PBC), Huberts Gamma(HG), Average Silhouette width (ASW), R2 (R2) , and Huberts Coefficient (HC) (Table 5.4.1). The clustering solution from data refined by SVM had the highest score for PBC and HG as well as the lowest score for HC. The solution from doubly refined data had the highest score for ASW and R2. For all scores, the unrefined data did the poorest, that is, it had the lowest score for PBC, HG, ASW, and R2 and the highest score for HC. According to all measures, refined data led to a solution that outperformed that from raw data.

Table 5.4.1. Cluster Quality Measures

Quality Measure	GDB	GDB_k	GDB_s	GDB_{ks}
Point Biserial Correlation	0.329	0.331	**0.392**	0.377
Hubert's Gamma	0.877	0.879	**0.932**	0.925
Average Silhouette width	0.146	0.138	0.295	**0.199**
R2	0.723	0.706	0.776	**0.785**
Hubert's Coefficient	0.0818	0.08377	**0.0556**	0.0589

5.4.5 Biological Measures

Cluster solutions produced from refined data sets were ranked as biologically more useful by measures of cluster composition. On average, clusters of GDBk had more in-group members than clusters of GDB. Further, clusters of GDBs had fewer out-of-group members than clusters of GDB. Finally, clusters form the doubly filtered data set GDBks had both more in-group members and fewer out-of-group members than clusters from the unfiltered data set GDB.

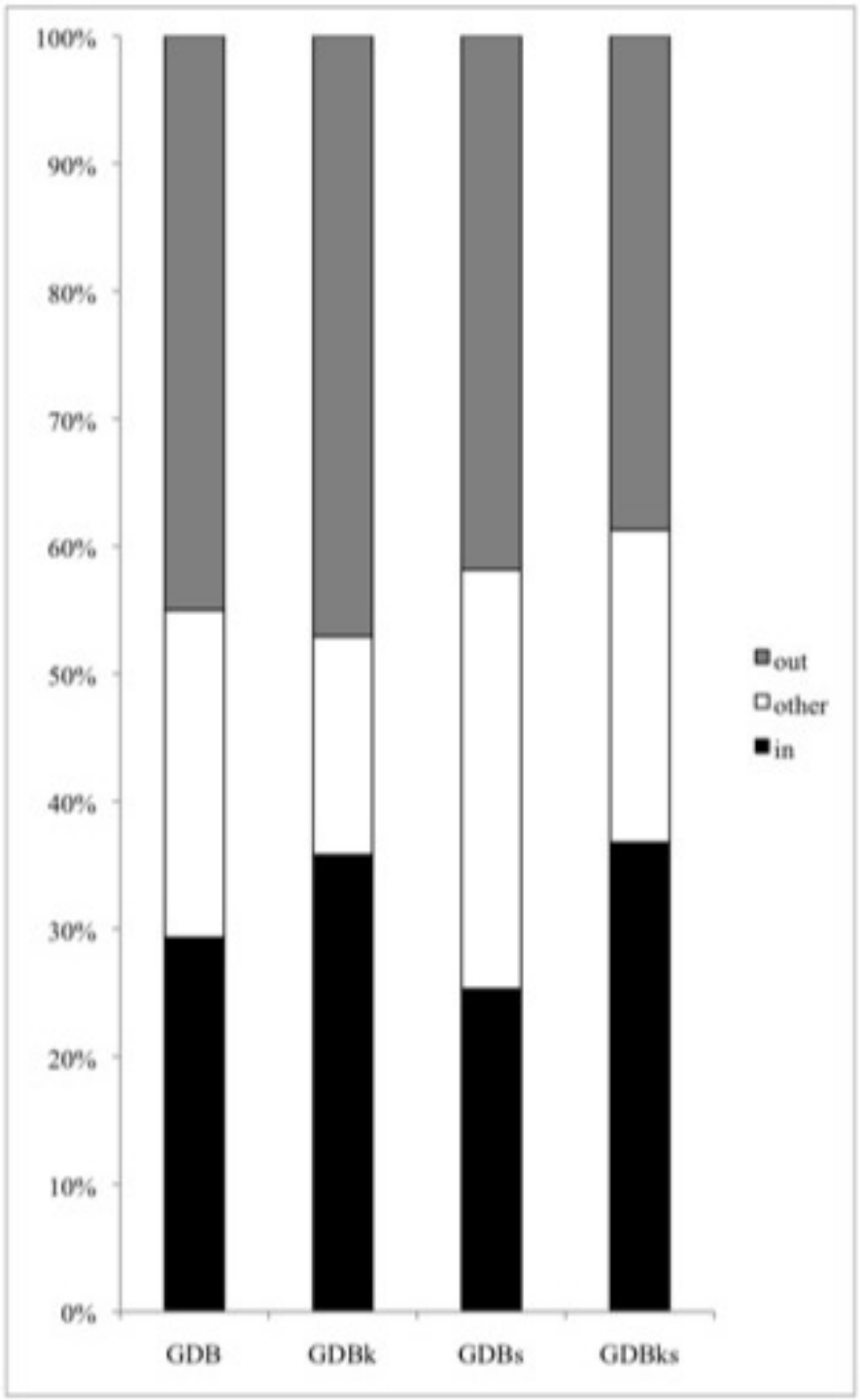

Figure 5.4.6. The average composition of random samples of 24 elements from each cluster with respect to in phenotypes, out phenotypes and other (images with numbers, plating artifact or missing colonies).

5.5 SVM

A two class SVM that was trained on images of WT BJB312 and mutants of the biofilm-defect cluster detected 276 mutants as biofilm deficient. The performance of this machine can be evaluated by its cross validation accuracy of 99.7%. Further, its performance is demonstrated when it is applied to a set of manually labeled data. Out of a set of 100 samples, it correctly identified 15 as biofilm mutants and 59 as other and it incorrectly labeled 8 as biofilm mutants and 18 as other.

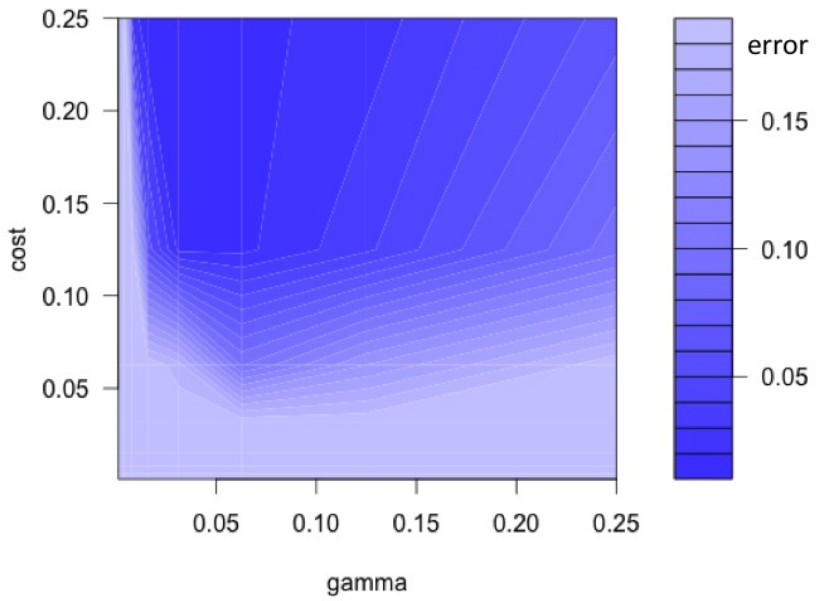

Figure 5.5.1. Grid search results for a two-class libSVM with a linear kernel trained on cluster F22 (size = 55) and BJB312 (size = 283) revealed the optimal parameters to be cost = 1 and gamma = 0.03125 achieving performance with error as low as 0.00344 and cross validation accuracy = 99.6

5.6 Functional Assays on Biofilm Mutants

Five of the ten mutants displayed lesser of biofilm/pellet in static liquid cultures of at least one media LB, TGHL, 1% Tryptone or R2A. One mutant showed a hyper

biofilm. Some of the five also showed corresponding decrease in violacein production in shaking liquid cultures. One mutant showed increase violacein production. Five of the same mutants exhibited smooth biofilms when plated on LB and TGHL. All strains, including BJB312, were smooth on 1% Tryptone and had minimal growth and violacein production on R2A. One mutant showed hyper wrinkled on 1% Tryptone. Three mutants exhibited abnormal swimming motility than BJB312 on LB. One mutant, shows a lack of violacein production in TGHL. Some of these mutants show less swarming motility, but it is not very dramatic due to BJB312s lack of swarming activity. In summary, the mutants differed noticeably from BJB312 in terms of their profile across medias in biofilm, pellet and violacein-production assays. There was no clear loss-of-function trends in the mutants with respect to swimming or swarming motility (Figureřefbehaviors).

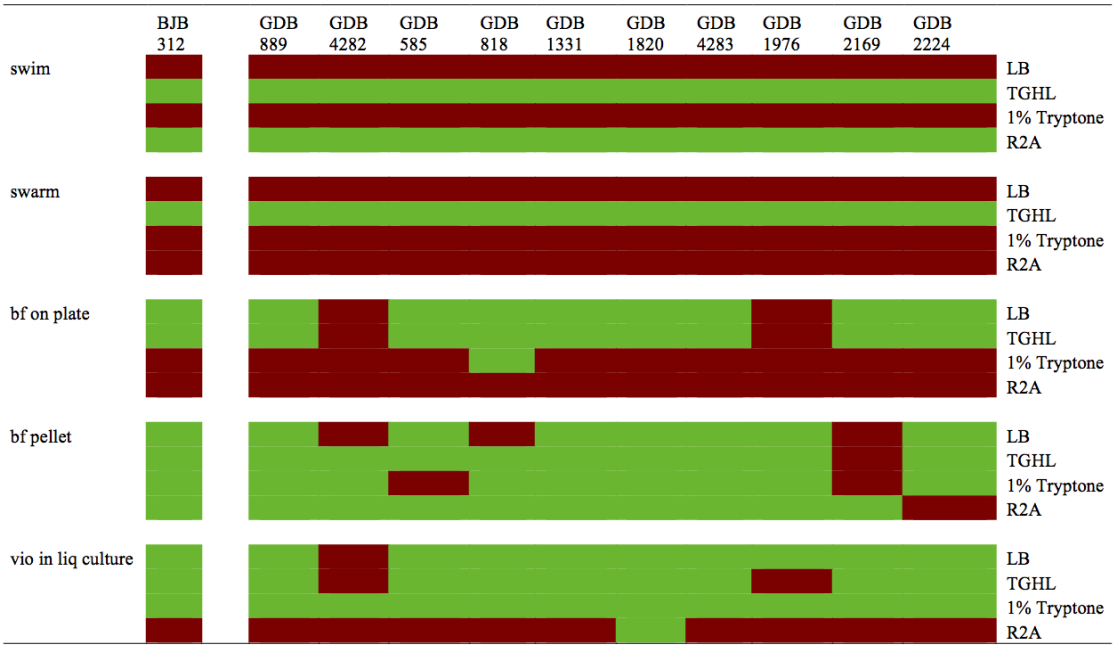

Figure 5.6.1. Representation of the behaviors of BJB312 and 10 insertion mutants across four media types (LB, TGHL, 1% Tryptone and R2A) with respect to (A) swimming motility, (B) swarming motility, (C) biofilm formation on a solid agar plates, (D) biofilm pellet formation in static liquid culture and (E) violacein production in agitated liquid culture. Green indicates a noticeable behavior and red indicates a lack of behavior.

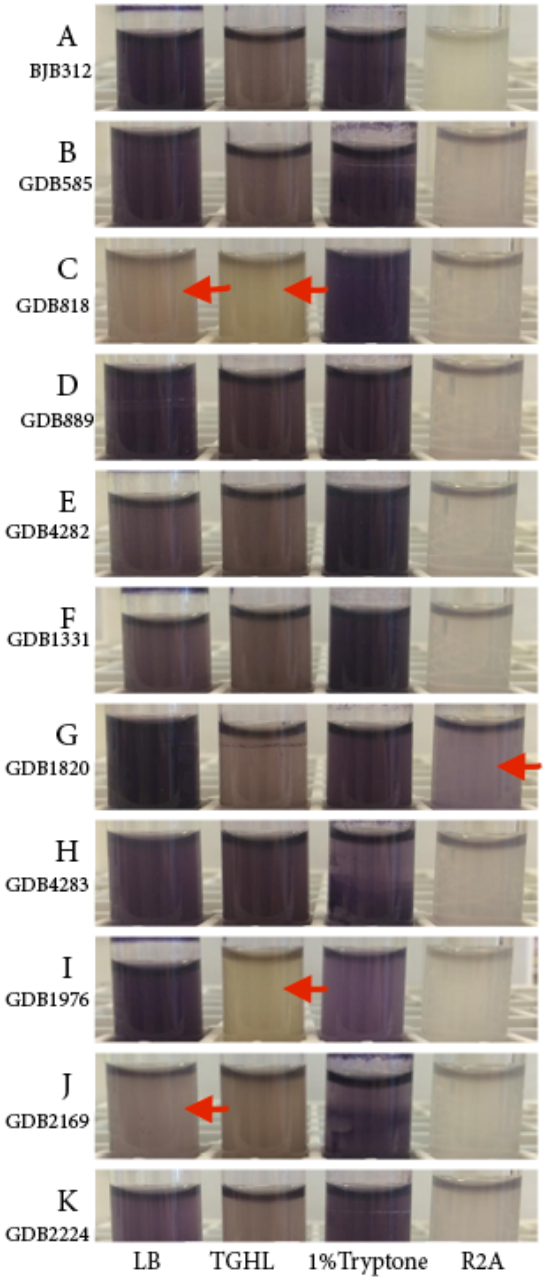

Figure 5.6.2. Environmental isolate BJB312 (A) and 10 transposon mutants (B-K) show differential violacein production when grown for 48 hours in shaking liquid medias of LB, TGHL, 1% Tryptone and R2A medias, left to right. Red arrows indicate abnormal phenotype.

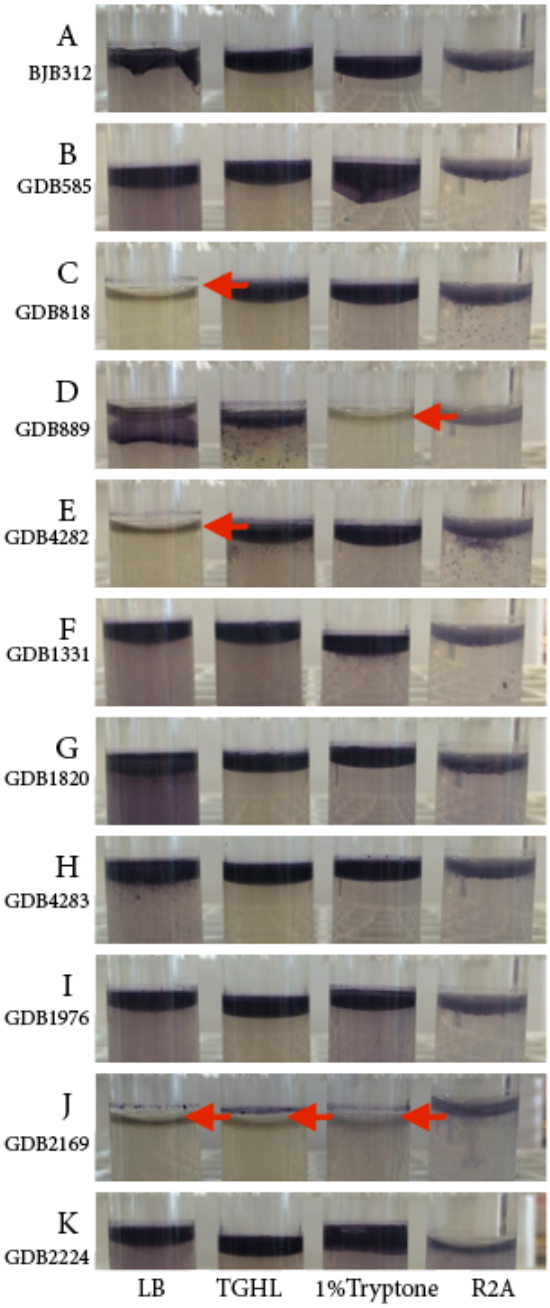

Figure 5.6.3. Environmental isolate BJB312 (A) and 10 transposon mutants (B-K) show differential biofilm pellet formation grown statically for 48 hours in liquid media of LB, TGHL, 1% Tryptone and R2A medias, left to right. Red arrows indicate abnormal phenotype.

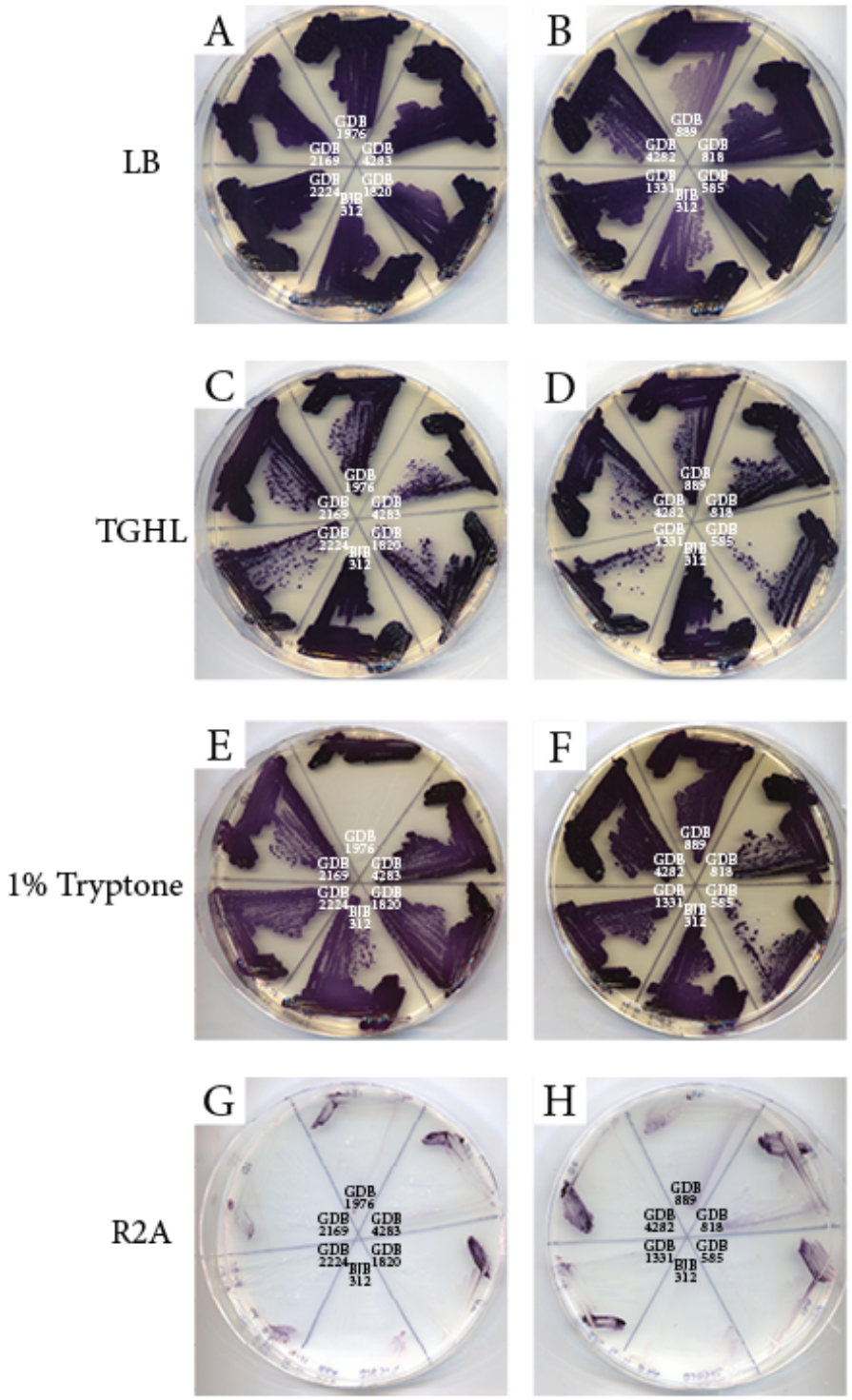

Figure 5.6.4. Environmental isolate BJB312 and 10 transposon mutants grown on 1.5% agar plates of the medias LB (A-B), TGHL (C-D), 1% Tryptone (E-F) and R2A (G-H). Contrasted against a white background.

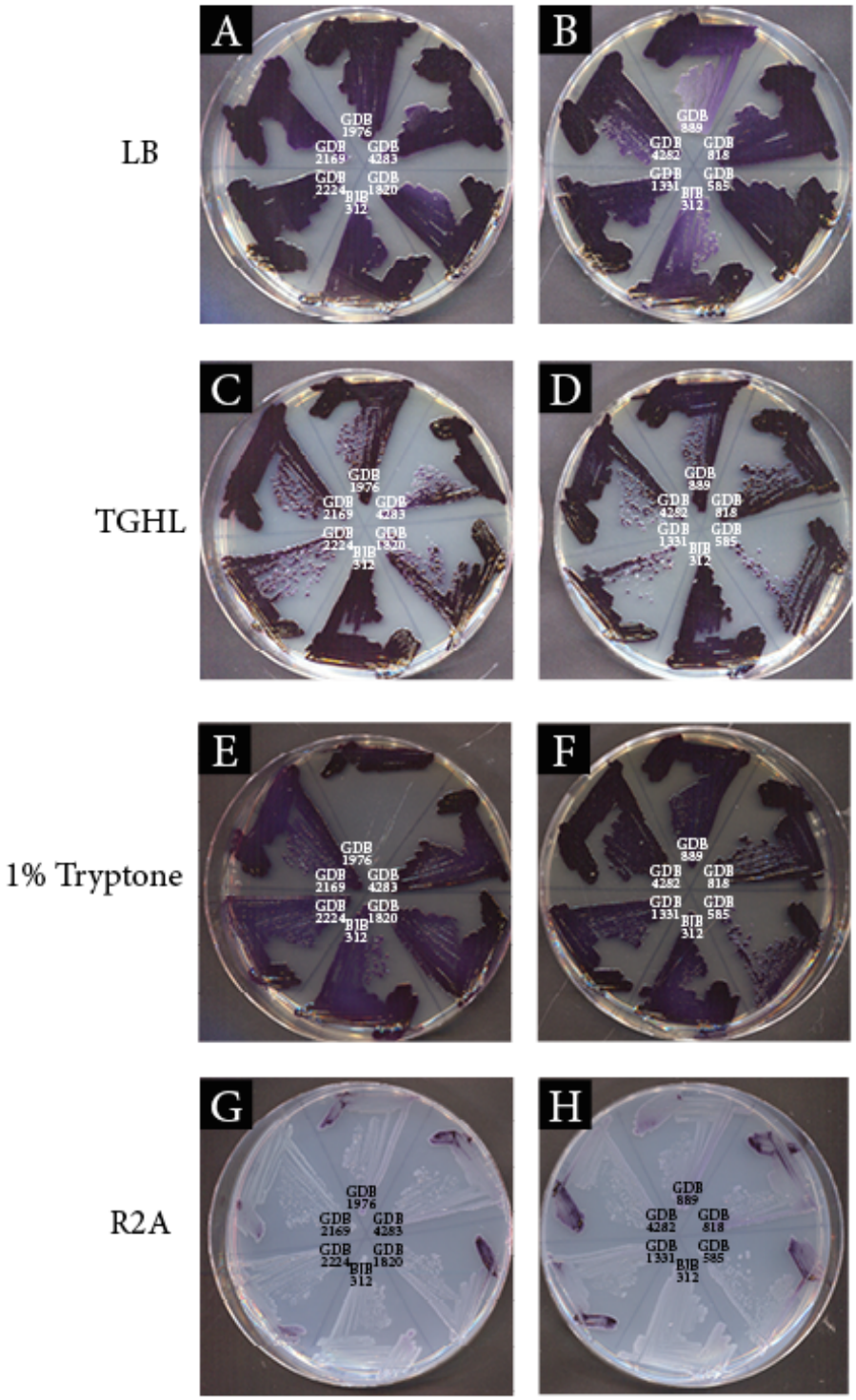

Figure 5.6.5. Environmental isolate BJB312 and 10 transposon mutants grown on 1.5% agar plates of the medias LB (A-B), TGHL (C-D), 1% Tryptone (E-F) and R2A (G-H). Contrasted against a black background to show subtle details.

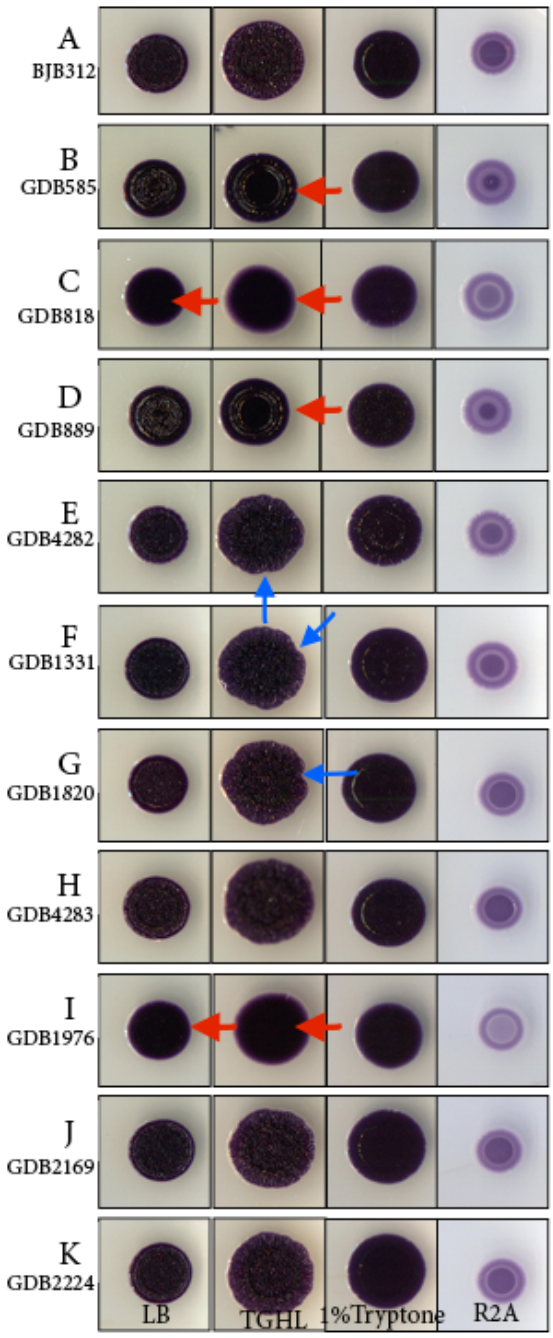

Figure 5.6.6. Environmental isolate BJB312 (A) and 10 transposon mutants (B-K) show differential biofilm formation grown on 1.5%agar plates of LB, TGHL, 1% Tryptone and R2A medias, left to right. Blue arrows indicate abnormal convexity phenotype. Red arrows indicate abnormal biofilm phenotype. Contrasted against a white background.

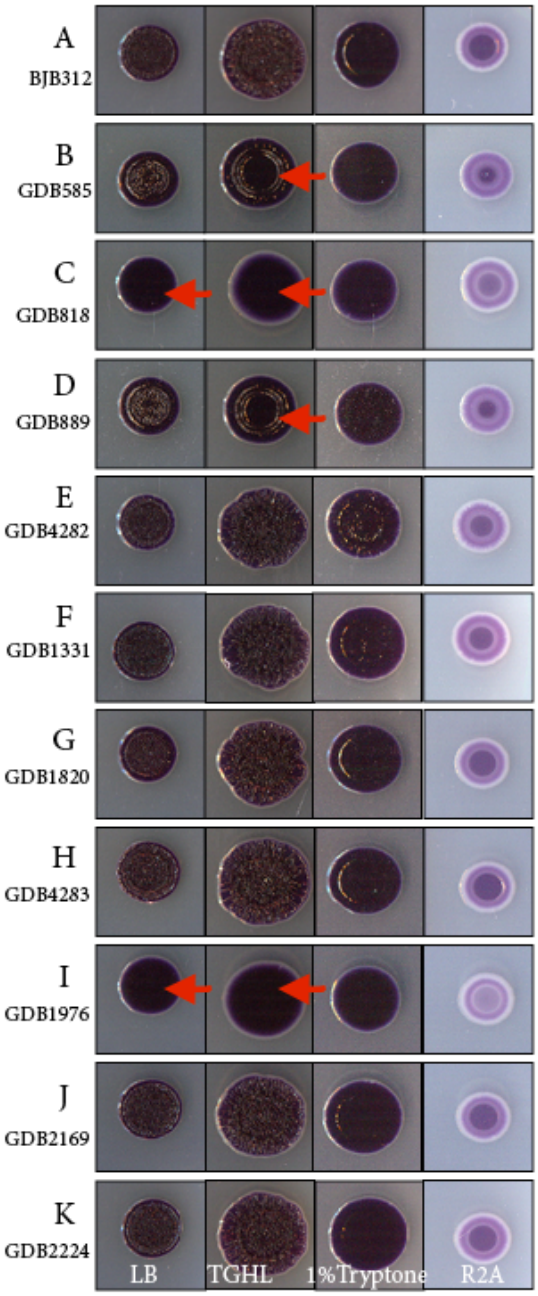

Figure 5.6.7. Environmental isolate BJB312 (A) and 10 transposon mutants (B-K) show differential biofilm formation grown on 1.5%agar plates of LB, TGHL, 1% Tryptone and R2A medias, left to right. Red arrows indicate abnormal phenotype. Contrasted against a black background to show subtle details.

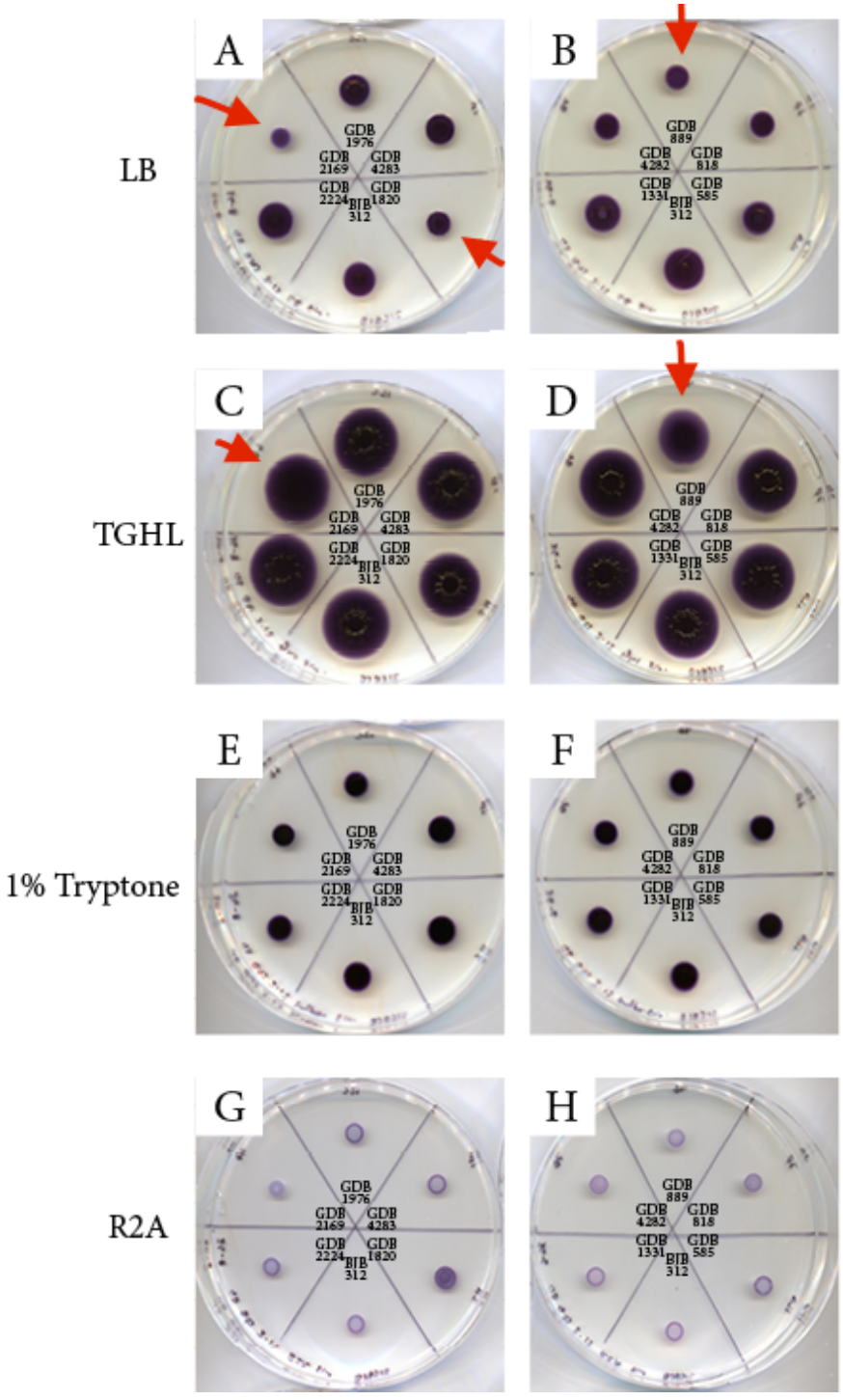

Figure 5.6.8. Environmental isolate BJB312 and 10 transposon mutants show differential swarming motility phenotypes in the medias (A-B) LB, (C-D)TGHL, (E-F) 1% Tryptone and (G-H) R2A. Red arrows indicate abnormal phenotype. Contrasted against a white background.

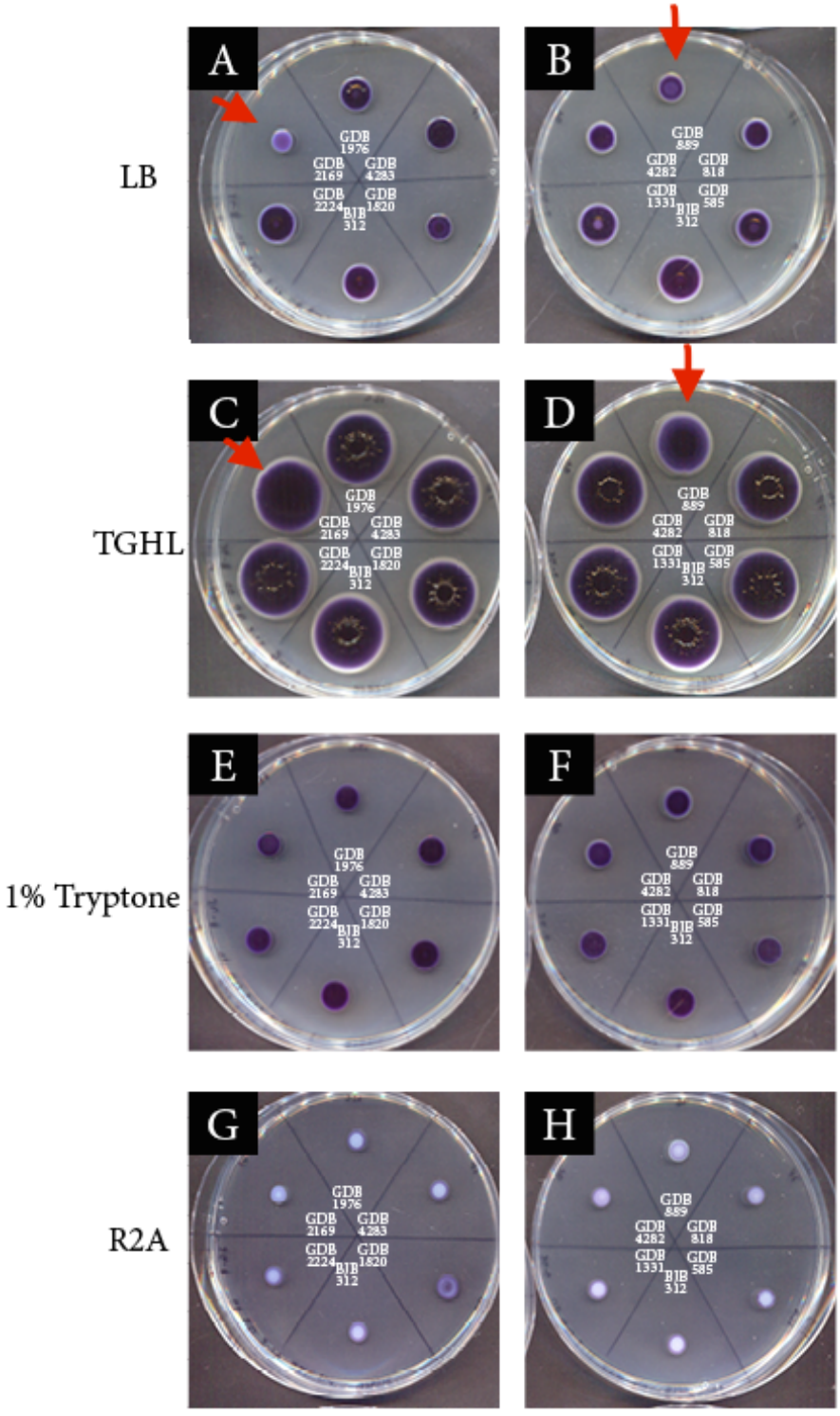

Figure 5.6.9. Environmental isolate BJB312 and 10 transposon mutants show differential swarming motility phenotypes in the medias (A-B) LB, (C-D)TGHL, (E-F) 1% Tryptone and (G-H) R2A. Red arrows indicate abnormal phenotype. Contrasted against a black background to show subtle details.

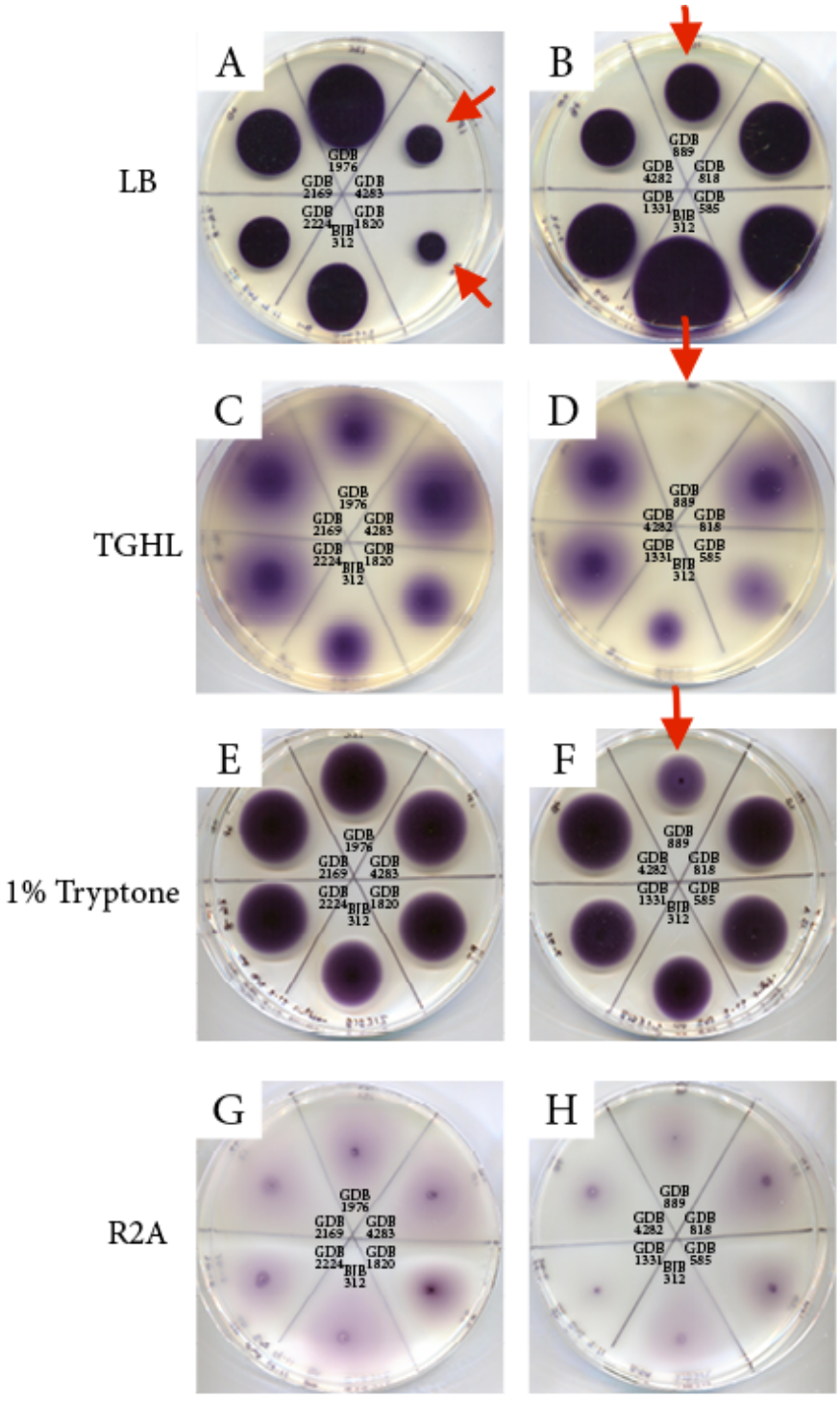

Figure 5.6.10. Environmental isolate BJB312 and 10 transposon mutants show differential swimming motility phenotypes in the medias (A-B) LB, (C-D)TGHL, (E-F) 1% Tryptone and (G-H) R2A. Red arrows indicate abnormal phenotype. Contrasted against a white background.

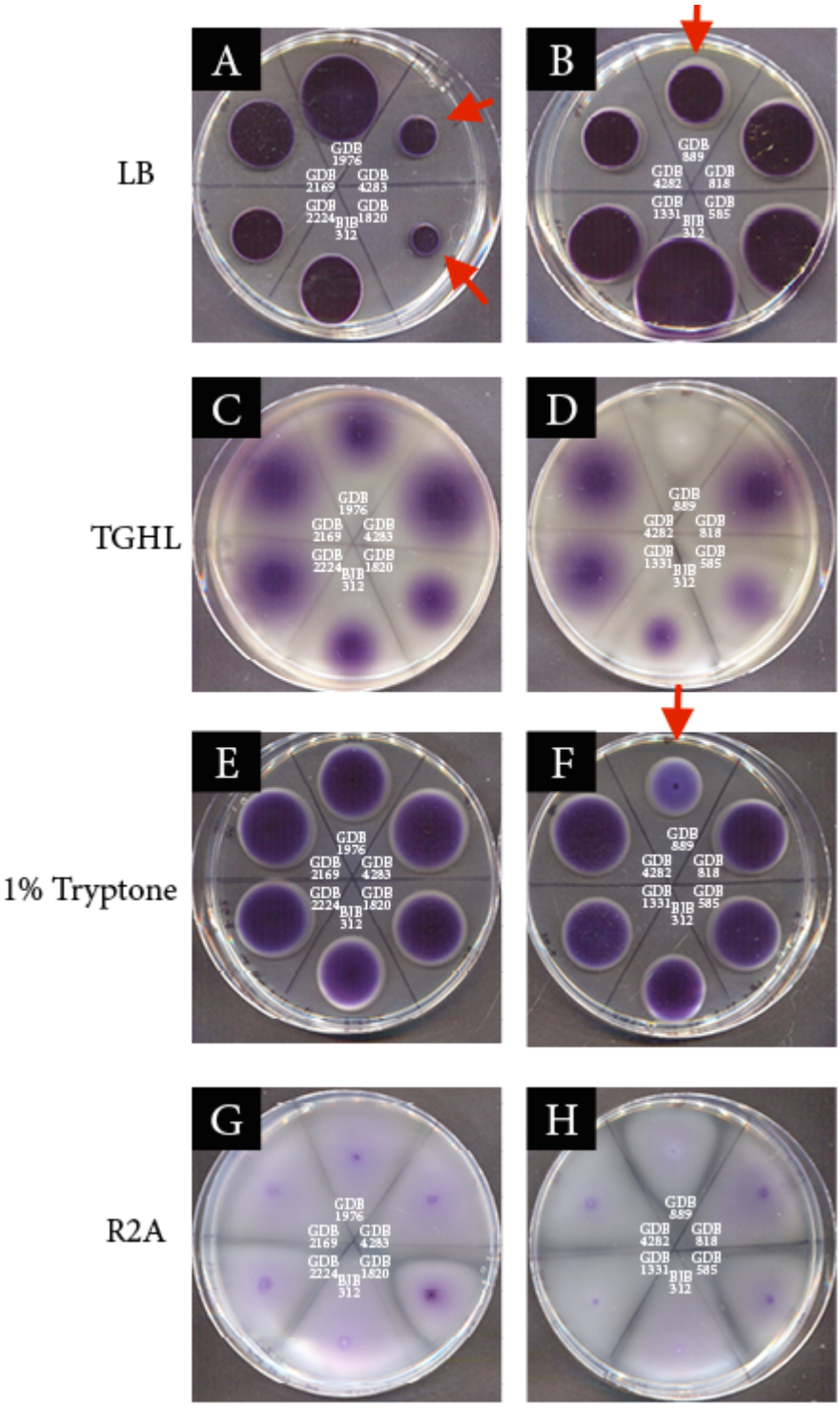

Figure 5.6.11. Environmental isolate BJB312 and 10 transposon mutants show differential swimming motility phenotypes in the medias (A-B) LB, (C-D)TGHL, (E-F) 1% Tryptone and (G-H) R2A. Red arrows indicate abnormal phenotype. Contrasted against a black background to show subtle details.

6
Discussion

6.1 Genomics

According to RAST, the nearest neighbor of BJB312 was *J. marseille*. This was surprising because *J. marseille* does not produce violacein. It is likely the case that there is no well documented genome of a *J. lividum* strain in the database used by RAST, but that BJB312 is in fact closer to *J. lividum*. It would be interesting to compare the genome of BJB312 to a range of *j. lividum* strains including available genomes such as HH01 and environmental isolates like BJB1. HH01 has been the subject of previous investigations in to *J. lividum* and could provide direction and comparative tools for the exploration of BJB312 [Hornung et al., 2013]. As violacein becomes more studies, there will likely need to be a more intricate taxonomy for these specific types of bacteria.

The genes of interest that were found int he BJB312 genome were related to EPS, cdGMP, quorum sensing, flagella and T2SS. Although these are different systems, they all interact with each other. Biofilm and motility are reciprocally regulated, likely by the cdGMP pathways [Lee et al., 2007]. Quorum sensing also seems to playa role in the program of violacein production and biofilm formation [Stauff and Bassler, 2011,Sakuragi

and Kolter, 2007]. Perhaps a future study on this library could use different features of images, or different assays, to keep track of phenotypes related to these different systems to study their interaction.

6.2 Mutant Library

Hand plating led to artifacts and unusable colonies. For example, one of the mating had nearly half white mutants. This is most likely contamination. As the library was built by hand, this isnt too unexpected. We could have automated this screen with a robot. More realistically, we could have maintained the library on a BJB312-specific antibiotic, like ampicillin. However, growth on antibiotics may affect the phenotype of the bacteria and select for a phenotype that is specifically resistant. We also could have done the matings in a hood. In general, batches with contamination should be excluded from the study, but because image processing didnt capture color-less colonies anyway, the contaminants didnt make it into the image library.

I combined two mutant libraries. They were of the same strain and the same transposon, but were plated by different people and scanned at different ages. Additionally, there were done on different media. This might be something to control for, but also shows the utility of the tool. It can be used to go back over a library, or augment a library. New questions can be asked about the same image library in that different phenotypic groups can be focused on, not just biofilm. New assays can be done and screened with ease and speed, for example motility or media, QS, response to salt etc.

Ultimately, my library could have been larger and more controlled. Given that BJB312 has about 5,000 genes in its genome, a very rough estimate is that I would need around 57,000 mutants to approach saturation [Phogot et al., 2001]. However, the machine learning methods I used seem more successful in light of the noisiness of my data; if they were applied to cleaner data I can imagine their performance increasing.

6.3 Biofilm Cluster

The mutants that were subjected to extensive functional assays were not those from the reported most successful cluster. They were from a cluster gathered in more preliminary clustering. Im not sure if this should be in the discussion, but it should be noted. Clearly, a further step in this experiment would be to examine the mutants from the most successful clusters as well as those classified as potentially similar. They would be subjected to functional assays, for example liquid biofilm assays, QS, temperature preference, and then sequenced. The next step in molecular biology would be clean deletions of genes of interest that were revealed though this too [Pelicic et al., 1997]. This is a more sophisticated way to determine gene function and ultimately allows us to modify isolates to combat chytrid better.

6.4 Image Processing

6.4.1 Image Segmentation

The method of image segmentation from far from ideal. Because I hand tuned the Hough Circle parameters to specify a radius range the a range of circles per image, it worked OK. However, there was low yield and certainly false positives. Better segmentation techniques, like the watershed method, could have improved the accuracy of image segmentation and collected a cleaner more complete image library.

6.4.2 Image Features

There have been extensive comparisons of image features and descriptors and their usefulness. Most of the features I chose provided usable information, but some (which ones) either returned a uniform measure or an incomplete data set. Although I looked at clustering with different combinations of the mutant classes, I could have looked at combinations of the ungrouped features. Using a weighted k-means approach did not

appear to offer any dramatic difference in clustering solutions, but weighting is certainly something I could have done and tried more sophisticated methods.

6.5 Machine Learning Filters

6.5.1 K-means filter vs SVM filter

It was interesting to note that the two filtering methods appeared to work qualitatively differently. The k-means based filter seemed to remove outliers, as illustrated by PCA. Further, it increased the average number of 'in' members in each cluster, but also increased the number of 'out' members. On the other hand, the SVM-based filter did not clearly remove any outliers, but rather thinned out the data to remove dense aggregates. Also, it reduced the average number of 'out' members per cluster but also decreased the number of 'in' members. When used together, the filters provided both advantages of increasing the number of 'in' members and decreasing the number of 'out' members compared to those of GDB. This relationship illustrates the qualitative differences and complimentary utilities of the two filtering methods.

6.5.2 One-class SVM

The one-class SVM was hyper sensitive resulting in a lot of false positive classifications. This resulted in a very dramatic filtration of the data set. Perhaps there are better ways to tune the parameters than adjust them so that is classifies approximately half of the training data as positive. Perhaps a larger WT library would allow for better training.

6.6 Comparing Clustering Solutions

I used statistical metrics to assess my clustering solutions. I also used subjective measures of biological usefulness. A more complete study would rescue the transposons from my class of biofilm mutants and sequence them to find out in which genes they have

inserted. At first a proof=of-principle study could look for known biofilm-related genes and this could validate the computational tools. Later, the tools could be trusted to find new, unexpected yet relevant genes.

6.7 Further work

The automatic screening tool could be used on other organisms, specifically those with biomedical significance such as cystic fibrosis colonizers. Biofilm plays a large role in cystic fibrosis and so it would be natural to extend this tool to look at biofilms in that species. Alternatively, it could be used to explore Gram positive bacteria. It could even be used to classify a collection of environmental isolates, they don't need to be mutants, just things that we want to find patterns in. This took provides a way to rapidly examine the phenotypes of a large library of bacteria and begin to make sense of the phenotypic groups present; it could be used with nearly any microbe and assay.

7
Conclusion

In this study I explored the potential of using image processing and Machine Learning to automatically screen large libraries of mutant bacterial colonies. These automation methods and this biological challenge seemed like a good pair because the former has been developed to find patterns in large complex data sets and the later provides exactly such a data set. Further, the question of how to scan large libraries in the biology laboratory is important because such data sets are currently the bottle-neck in high throughput experiments.

I built a library of ∼4,000 transposon mutants from the environmental isolate BJB312. Although this library did not come close to providing saturation of the genome, it is still the largest library of BJB312 made to date and sufficed as a set of data on which I could develop automated screening techniques.

Because I was dealing with a real-world data set, the questions of clustering and classification were not straightforward. I used unsupervised and supervised machine learning techniques to try and extract a phenotypically similar, biologically interesting and comprehensive group of mutants from the library. The phenotype of interest turned out to be

biofilm formation; biofilm was observable in image of the colonies and has important roles in the characteristic survival and therapeutic potentials of BJB312.

Time prohibited me from re-examine my final group of 276 biofilm mutants with extensive functional assays or genetic sequencing. However, I was able to explore 10 members of a group of biofilm mutants from a preliminary clustering and classification workflow. These members displayed differential abnormal phenotypes in assays of biofilm growth on solid media, pellet formation in liquid media and violacein production in liquid media. They did not show noticeable motility defects. This study leaves room for much more exploration or engineering a machine learning solution for characterizing images of microbes. However, these preliminary results suggest that this workflow may prove to be a useful tool.

References

[Agematu et al., 2011] Agematu, H., Suzuki, K., and Tsuya, H. (2011). Massilia sp. BS-1, a Novel Violacein-Producing Bacterium Isolated from Soil. *Bioscience, Biotechnology, and Biochemistry*, 75(10):2008–2010.

[Alavi and Belas, 2001] Alavi, M. and Belas, R. (2001). [3] Surface sensing, swarmer cell differentiation, and biofilm development. In *sciencedirect.com.ezprox.bard.edu*, pages 29–40. Elsevier.

[Alrabea et al., 2013] Alrabea, A., Senthilkumar, A. V., Al-Shalabi, H., and Bader, A. (2013). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA. *Journal of Advances in Computer Networks*, pages 137–142.

[Andrighetti-Frohner et al., 2003] Andrighetti-Frohner, C. R., Antonio, R. V., Creczynski-Pasa, T. B., Barardi, C., and Simoes, C. (2003). Cytotoxicity and potential antiviral evaluation of violacein produced by Chromobacterium violaceum. *Memórias do Instituto Oswaldo Cruz*, 98(6):843–848.

[Andrighetti-Fröhner et al., 2006] Andrighetti-Fröhner, C. R., Kratz, J. M., Antonio, R. V., Creczynski-Pasa, T. B., Barardi, C. R. M., and Simões, C. M. O. (2006). In vitro testing for genotoxicity of violacein assessed by Comet and Micronucleus assays. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 603(1):97–103.

[Aranda et al., 2011] Aranda, S., Montes-Borrego, M., and Landa, B. B. (2011). Purple-pigmented violacein-producing Duganella spp. inhabit the rhizosphere of wild and cultivated olives in southern Spain. *Microbial ecology*, 62(2):446–459.

[August et al., 2000] August, P. R., Grossman, T. H., Minor, C., Draper, M. P., MacNeil, I. A., Pemberton, J. M., Call, K. M., Holt, D., and Osburne, M. S. (2000). Sequence Analysis and Functional Characterization of the Violacein Biosynthetic Pathway from Chromobacterium violaceum. *Journal of Molecular Biology*, 2(4):513–519.

[Autret and Charbit, 2005] Autret, N. and Charbit, A. (2005). Lessons from signature-tagged mutagenesis on the infectious mechanisms of pathogenic bacteria. *FEMS Microbiology Reviews*, 29(4):703–717.

[Becker et al., 2009] Becker, M. H., Brucker, R. M., Schwantes, C. R., Harris, R. N., and Minbiole, K. P. C. (2009). The Bacterially Produced Metabolite Violacein Is Associated with Survival of Amphibians Infected with a Lethal Fungus. *Applied and Environmental Microbiology*, 75(21):6635–6638.

[Becker and Harris, 2010] Becker, M. H. and Harris, R. N. (2010). Cutaneous Bacteria of the Redback Salamander Prevent Morbidity Associated with a Lethal Disease. *PLoS ONE*, 5(6):e10957.

[Beghdadi et al., 2013] Beghdadi, A., Larabi, M. C., Bouzerdoum, A., and Iftekharuddin, K. M. (2013). A survey of perceptual image processing methods. *Signal Processing: Image Communication*, 28(8):811–831.

[Belden and Harris, 2007] Belden, L. K. and Harris, R. N. (2007). Infectious diseases in wildlife: the community ecology context. *Frontiers in Ecology and the Environment*, 5(10):533–539.

[Ben-Hur and Weston, 2009] Ben-Hur, A. and Weston, J. (2009). A User's Guide to Support Vector Machines. In *Data mining techniques for the life sciences*, pages 223–239. Humana Press, Totowa, NJ.

[Berger et al.,] Berger, L., Speare, R., Daszak, P., Green, D. E., Cunningham, A. A., Goggin, C. L., Slocombe, R., Ragan, M. A., Hyatt, A. D., McDonald, K. R., Hines, H. B., Lips, K. R., Marantelli, G., and Parkes, H. Chytridiomycosis causes amphibian mortality associated with population declines in the rain forests of Australia and Central America. *pnas.org*.

[Biggs and hickey, 1994] Biggs, B. J. F. and hickey, c. w. (1994). Periphyton responses to a hydraulic gradient in a regulated river in New Zealand. *Freshwater Biology*, 32(1):49–59.

[Bjarnsholt et al., 2005] Bjarnsholt, T., Jensen, P. Ø., Burm∘$lle, M., Hentzer, M., Haagensen, J.A.\ J., Hougen, H.\ P., Calum, H., Madsen, K.\ G., Moser, C., M$
$373 - -383$.

[Bletz et al., 2013] Bletz, M. C., Loudon, A. H., Becker, M. H., Bell, S. C., Woodhams, D. C., Minbiole, K. P. C., and Harris, R. N. (2013). Mitigating amphibian chytridiomycosis with bioaugmentation: characteristics of effective probiotics and strategies for their selection and use. *Ecology Letters*, 16(6):807–820.

[Brucker et al., 2008] Brucker, R. M., Harris, R. N., Schwantes, C. R., Gallaher, T. N., Flaherty, D. C., Lam, B. A., and Minbiole, K. P. C. (2008). Amphibian chemical

defense: antifungal metabolites of the microsymbiont Janthinobacterium lividum on the salamander Plethodon cinereus. *Journal of chemical ecology*, 34(11):1422–1429.

[Bubeck et al., 2013] Bubeck, S. e., Meila, M., and von Luxburg, U. (2013). HOW THE INITIALIZATION AFFECTS THE STABILITY OF THE k-MEANS ALGORITHM ∗ . In *ESAIM Probability and Statsitics*.

[Carey, 1993] Carey, C. (1993). Hypothesis Concerning the Causes of the Disappearance of Boreal Toads from the Mountains of Colorado. *Conservation Biology*, 7(2):355–362.

[Carpentier and Cerf, 1993] Carpentier, B. and Cerf, O. (1993). Biofilms and their consequences, with particular reference to hygiene in the food industry. *The Journal of applied bacteriology*, 75(6):499–511.

[Celebi et al., 2013] Celebi, M. E., Kingravi, H. A., and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, 40(1):200–210.

[Chiang and Mekalanos, 1998] Chiang, S. L. and Mekalanos, J. J. (1998). Use of signature-tagged transposon mutagenesis to identify Vibrio choleraegenes critical for colonization. *Molecular Microbiology*, 27(4):797–805.

[Clausi, 2002] Clausi, D. A. (2002). K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation. *Pattern Recognition*, 35(9):1959–1972.

[Coquet et al.,] Coquet, L., Junter, G. A., and Jouenne, T. Resistance of artificial biofilms of Pseudomonas aeruginosa to imipenem and tobramycin.

[Cordeiro de Amorim and Mirkin, 2012] Cordeiro de Amorim, R. and Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recognition*, 45(3):1061–1075.

[Corn and Fogleman, 1984] Corn, P. S. and Fogleman, J. C. (1984). Extinction of Montane Populations of the Northern Leopard Frog (Rana pipiens) in Colorado. *Journal of Herpetology*, 18(2):147.

[Costerton et al., 1987] Costerton, J. W., Cheng, K. J., Geesey, G. G., Ladd, T. I., Nickel, J. C., Dasgupta, M., and Marrie, T. J. (1987). Bacterial biofilms in nature and disease. *Annual Review of Microbiology*, 41:435–464.

[Davey et al., 2003] Davey, M. E., Caiazza, N. C., and O'Toole, G. A. (2003). Rhamnolipid Surfactant Production Affects Biofilm Architecture in Pseudomonas aeruginosa PAO1. *Journal of Bacteriology*, 185(3):1027–1036.

[Davies et al., 1998] Davies, D. G., Parsek, M. R., Pearson, J. P., Iglewski, B. H., Costerton, J. W., and Greenberg, E. P. (1998). The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science*, 280(5361):295–298.

[de Lorenzo and Timmis, 1994] de Lorenzo, V. and Timmis, K. N. (1994). Analysis and construction of stable phenotypes in gram-negative bacteria with Tn5- and Tn10-derived minitransposons. *Methods in enzymology*, 235:386–405.

[Di Lazzaro et al., 2013] Di Lazzaro, P., Murra, D., and Schwortz, B. (2013). Pattern recognition after image processing of low-contrast images, the case of the Shroud of Turin. *Pattern Recognition*, 46(7):1964–1970.

[Duran et al., 2007] Duran, N., Justo, G. Z., Ferreira, C. V., Melo, P. S., Cordi, L., and Martins, D. (2007). Violacein: properties and biological activities. *Biotechnology and Applied Biochemistry*, 48(3):127.

[Duran and Menck, 2001] Duran, N. and Menck, C. F. M. (2001). Chromobacterium violaceum: A Review of Pharmacological and Industiral Perspectives. *Critical Reviews in Microbiology*, 27(3):201–222.

[El Agha and M Ashour, 2012] El Agha, M. and M Ashour, W. (2012). Efficient and Fast Initialization Algorithm for K-means Clustering. *International Journal of Intelligent Systems and Applications*, 4(1):21–31.

[Eng et al., 1991] Eng, R. H., Padberg, F. T., Smith, S. M., Tan, E. N., and Cherubiame, C. E. (1991). Bactericidal effects of antibiotics on slowly growing and nongrowing bacteria. *Antimicrobial Agents and Chemotherapy*, 35(9):1824.

[Fan et al., 2011] Fan, Y., Zhang, S., Kruer, N., and Keyhani, N. O. (2011). High-throughput insertion mutagenesis and functional screening in the entomopathogenic fungus Beauveria bassiana. *Journal of Invertebrate Pathology*, 106(2):274–279.

[Ferreira, 2004] Ferreira, C. V. (2004). Molecular mechanism of violacein-mediated human leukemia cell death. *Blood*, 104(5):1459–1464.

[Foulongne et al., 2000] Foulongne, V., Bourg, G., Cazevieille, C., Michaux-Charachon, S., and O'Callaghan, D. (2000). Identification of Brucella suis Genes Affecting Intracellular Survival in an In Vitro Human Macrophage Infection Model by Signature-Tagged Transposon Mutagenesis. *Infection and Immunity*, 68(3):1297–1303.

[Friedman and Kolter, 2004] Friedman, L. and Kolter, R. (2004). Genes involved in matrix formation in Pseudomonas aeruginosa PA14 biofilms. *Molecular Microbiology*, 51(3):675–690.

[Gidudu Anthony and Tshilidzi,] Gidudu Anthony, H. G. and Tshilidzi, M. Image Classification Using SVMs: One-against-One Vs One-against-All .

[Graba et al., 2013] Graba, M., Sauvage, S., Moulin, F. Y., Urrea, G., Sabater, S., and Sanchez-Pérez, J. M. (2013). Interaction between local hydrodynamics and algal community in epilithic biofilm. *Water Research*, 47(7):2153–2163.

[Gristina, 1987] Gristina, A. G. (1987). Biomaterial-centered infection: microbial adhesion versus tissue integration. *Science*, 237(4822):1588–1595.

[Guo et al., 2015] Guo, J., Liang, T., Hu, C., Lv, R., Yang, X., Cui, Y., Song, Y., Yang, R., Zhu, Q., and Song, Y. (2015). Sequence types diversity of Legionella pneumophila isolates from environmental water sources in Guangzhou and Jiangmen, China. *Infection, Genetics and Evolution*, 29:35–41.

[Haagensen et al., 2006] Haagensen, J. A. J., Klausen, M., Ernst, R. K., Miller, S. I., Folkesson, A., Tolker-Nielsen, T., and Molin, S. (2006). Differentiation and Distribution of Colistin- and Sodium Dodecyl Sulfate-Tolerant Cells in Pseudomonas aeruginosa Biofilms. *Journal of Bacteriology*, 189(1):28–37.

[Hakvåg et al., 2009] Hakvåg, S., Fjærvik, E., Klinkenberg, G., Borgos, S. E. F., Josefsen, K. D., Ellingsen, T. E., and Zotchev, S. B. (2009). Violacein-Producing Collimonas sp. from the Sea Surface Microlayer of Costal Waters in Tr$\circ ndelag, Norway. Marine Drugs, 7(4) : 576 - -588$.

[Harris et al., 2009] Harris, R. N., Lauer, A., Simon, M. A., Banning, J. L., and Alford, R. A. (2009). Addition of antifungal skin bacteria to salamanders ameliorates the effects of chytridiomycosis. *Diseases of Aquatic Organisms*, 83:11–16.

[Hentzer et al., 2001] Hentzer, M., Teitzel, G. M., Balzer, G. J., Heydorn, A., Molin, S., Givskov, M., and Parsek, M. R. (2001). Alginate Overproduction Affects Pseudomonas aeruginosa Biofilm Structure and Function. *Journal of Bacteriology*, 183(18):5395–5401.

[Hornung et al., 2013] Hornung, C., Poehlein, A., Haack, F. S., Schmidt, M., Dierking, K., Pohlen, A., Schulenburg, H., Blokesch, M., Plener, L., Jung, K., Bonge, A., Krohn-Molt, I., Utpatel, C., Timmermann, G., Spieck, E., Pommerening-Röser, A., Bode, E., Bode, H. B., Daniel, R., Schmeisser, C., and Streit, W. R. (2013). The Janthinobacterium sp. HH01 Genome Encodes a Homologue of the V. cholerae CqsA and L. pneumophila LqsA Autoinducer Synthases. *PLoS ONE*, 8(2):e55045.

[Hoshino, 2011] Hoshino, T. (2011). Violacein and related tryptophan metabolites produced by Chromobacterium violaceum: biosynthetic mechanism and pathway for construction of violacein core. *Applied Microbiology and Biotechnology*, 91(6):1463–1475.

[Hosseinpour et al., 2014] Hosseinpour, S., Rafiee, S., Aghbashlo, M., and Mohtasebi, S. S. (2014). A novel image processing approach for in-line monitoring of visual texture during shrimp drying. *Journal of Food Engineering*, 143:154–166.

[Hu et al., 2008] Hu, G., Zhou, S., Guan, J., and Hu, X. (2008). Towards effective document clustering: A constrained K-means based approach. *Information Processing & Management*, 44(4):1397–1409.

[Huang and Zeng, 2015] Huang, C. and Zeng, L. (2015). Robust image segmentation using local robust statistics and correntropy-based K-means clustering. *Optics and Lasers in Engineering*, 66:187–203.

[Huang et al., 2005] Huang, J. Z., Ng, M. K., Rong, H., and Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):657–668.

[Hutchison III, 1999] Hutchison III, C. A. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science*, 286(5447):2165–2169.

[Hyduke and Palsson, 2010] Hyduke, D. R. and Palsson, B. O. (2010). Towards genome-scale signalling network reconstructions. *Nature reviews. Genetics*, 11(4):297–307.

[Jackson et al., 2004] Jackson, K. D., Starkey, M., Kremer, S., Parsek, M. R., and Wozniak, D. J. (2004). Identification of psl, a Locus Encoding a Potential Exopolysaccharide That Is Essential for Pseudomonas aeruginosa PAO1 Biofilm Formation. *Journal of Bacteriology*, 186.

[Jacobs et al., 2011] Jacobs, M. A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., Chun-Rong, L., Guenthner, D., Bovee, D., Olson, M. V., and Manoil, C. (2011). Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24):14339–14344.

[Jeyapoovan and Murugan, 2013] Jeyapoovan, T. and Murugan, M. (2013). Surface roughness classification using image processing. *Measurement*, 46(7):2065–2072.

[Johnson and Speare, 2003] Johnson, M. L. and Speare, R. (2003). Survival of Batrachochytrium dendrobatidis in Water: Quarantineand Disease Control Implications. *Emerging Infectios Disease*, 9(8):1–4.

[Joneson et al., 2011] Joneson, S., Stajich, J. E., Shiu, S.-H., and Rosenblum, E. B. (2011). Genomic Transition to Pathogenicity in Chytrid Fungi. *PLoS Pathogens*, pages 1–11.

[Juang and Wu, 2011] Juang, L.-H. and Wu, M.-N. (2011). Psoriasis image identification using k-means clustering with morphological processing. *Measurement*, 44(5):895–905.

[Judson and Mekalanos, 2000] Judson, N. and Mekalanos, J. J. (2000). Transposon-based approaches to identify essential bacterial genes. *Trends in Microbiology*, 8(11):521–526.

[Kearns et al., 2001] Kearns, D. B., Robinson, J., and Shimkets, L. J. (2001). Pseudomonas aeruginosa Exhibits Directed Twitching Motility Up Phosphatidylethanolamine Gradients. *Journal of Bacteriology*, 183(2):763–767.

[Keesing et al., 2010] Keesing, F., Belden, L. K., Daszak, P., Dobson, A., Harvell, C. D., Holt, R. D., Hudson, P., Jolles, A., Jones, K. E., Mitchell, C. E., Myers, S. S., Bogich, T., and Ostfeld, R. S. (2010). Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature*, 468(7324):647–652.

[Kistner et al., 2013] Kistner, M., Jemwa, G. T., and Aldrich, C. (2013). Monitoring of mineral processing systems by using textural image analysis. *Minerals Engineering*, 52:169–177.

[Kodach et al., 2006] Kodach, L. L., Bos, C. L., Duran, N., Peppelenbosch, M. P., Ferreira, C. V., and Hardwick, J. C. H. (2006). Violacein synergistically increases 5-fluorouracil cytotoxicity, induces apoptosis and inhibits Akt-mediated signal transduction in human colorectal cancer cells. *Carcinogenesis*, 27(3):508–516.

[Kosmidis et al., 2012] Kosmidis, C., Schindler, B. D., Jacinto, P. L., Patel, D., Bains, K., Seo, S. M., and Kaatz, G. W. (2012). Expression of multidrug resistance efflux pump genes in clinical and environmental isolates of Staphylococcus aureus. *International Journal of Antimicrobial Agents*, 40(3):204–209.

[Lam et al., 2010] Lam, B. A., Walke, J. B., Vredenburg, V. T., and Harris, R. N. (2010). Proportion of individuals with anti-Batrachochytrium dendrobatidis skin bacteria is associated with population persistence in the frog Rana muscosa. *Biological Conservation*, 143(2):529–531.

[Lam et al., 2011] Lam, B. A., Walton, D. B., and Harris, R. N. (2011). Motile zoospores of Batrachochytrium dendrobatidis move away from antifungal metabolites produced by amphibian skin bacteria. *EcoHealth*, 8(1):36–45.

[Landell et al., 2013] Landell, M. F., Salton, J., Caumo, K., Broetto, L., and Rott, M. B. (2013). Isolation and genotyping of free-living environmental isolates of Acanthamoeba spp. from bromeliads in Southern Brazil. *Experimental Parasitology*, 134(3):290–294.

[Landry et al., 2006] Landry, R. M., An, D., Hupp, J. T., Singh, P. K., and Parsek, M. R. (2006). Mucin-Pseudomonas aeruginosa interactions promote biofilm formation and antibiotic resistance. *Molecular Microbiology*, 59(1):142–151.

[Larsen et al., 2002] Larsen, R., Wilson, M., Guss, A., and Metcalf, W. (2002). Genetic analysis of pigment biosynthesis in Xanthobacter autotrophicus Py2 using a new, highly efficient transposon mutagenesis system that is functional in a wide variety of bacteria. *Archives of Microbiology*, 178(3):193–201.

[Lauer et al., 2007a] Lauer, A., Simon, M.-A., Banning, J. L., André, E., Duncan, K., and Harris, R. N. (2007a). Common Cutaneous Bacteria from the Eastern Red-Backed Salamander Can Inhibit Pathogenic Fungi. *Copeia*, 2007:630–640.

[Lauer et al., 2007b] Lauer, A., Simon, M.-A., Banning, J. L., Lam, B. A., and Harris, R. N. (2007b). Diversity of cutaneous bacteria with antifungal activity isolated from female four-toed salamanders. *The ISME Journal*, 2(2):145–157.

[Lawrence et al., 1989] Lawrence, J. R., Korber, D. R., and Caldwell, D. E. (1989). Computer-enhanced darkfield microscopy for the quantitative analysis of bacterial growth and behavior on surfaces. *Journal of Microbiological Methods*, 10(2):123–138.

[Lee et al., 2006] Lee, C.-O., Jeon, K., Ha, Y., and Hahn, J. (2006). A variational approach to blending based on warping for non-overlapped images. *Computer Vision and Image Understanding*, (105):112–220.

[Lee et al., 2007] Lee, V. T., Matewish, J. M., Kessler, J. L., Hyodo, M., Hayakawa, Y., and Lory, S. (2007). A cyclic-di-GMP receptor required for bacterial exopolysaccharide production. *Molecular Microbiology*, 65(6):1474–1484.

[Lefevre and Watkins, 1986] Lefevre, G. and Watkins, W. (1986). The Question Of The Total Gene Number In Drosophila Melanogaster . *Genetics*, 113:869–895.

[Leiva and Vidal, 2013] Leiva, L. A. and Vidal, E. (2013). Warped K-Means: An algorithm to cluster sequentially-distributed data. *Information Sciences*, 237:196–210.

[Liao et al., 2013] Liao, K., Liu, G., Xiao, L., and Liu, C. (2013). A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval. *Knowledge-Based Systems*, 49:123–133.

[Longcore et al., 2014] Longcore, J., Pessier, A. P., and Nichols, D. K. (2014). Batrachochytriumdendrobatidisgen. et sp. nov., a chytrid pathogenic to amphibians. pages 1–10.

[Lopes et al., 2009] Lopes, S. C. P., Blanco, Y. C., Justo, G. Z., Nogueira, P. A., Rodrigues, F. L. S., Goelnitz, U., Wunderlich, G., Facchini, G., Brocchi, M., Duran, N., and Costa, F. T. M. (2009). Violacein Extracted from Chromobacterium violaceum Inhibits Plasmodium Growth In Vitro and In Vivo. *Antimicrobial Agents and Chemotherapy*, 53(5):2149–2152.

[Mah and O'Toole, 2001] Mah, T. F. and O'Toole, G. A. (2001). Mechanisms of biofilm resistance to antimicrobial agents. *Trends in Microbiology*, 9(1):34–39.

[Marshall et al., 1971] Marshall, K. C., Stout, R., and Mitchell, R. (1971). Mechanism of the Initial Events in the Sorption of Marine Bacteria to Surfaces. *Journal of General Microbiology*, 68(3):337–348.

[Matsukawa and Greenberg, 2004] Matsukawa, M. and Greenberg, E. P. (2004). Putative Exopolysaccharide Synthesis Genes Influence Pseudomonas aeruginosa Biofilm Development. *Journal of Bacteriology*, 186(14):4449–4456.

[Metcalf et al., 1996] Metcalf, W. W., Jiang, W., Daniels, L. L., Kim, S.-K., Haldimann, A., and Wanner, B. L. (1996). Conditionally Replicative and Conjugative Plasmids CarryinglacZα for Cloning, Mutagenesis, and Allele Replacement in Bacteria. *Plasmid*, 35(1):1–13.

[Meyer et al., 2012] Meyer, E. A., Cramp, R. L., Bernal, M. H., and Franklin, C. E. (2012). Changes in cutaneous microbial abundance with sloughing: possible implications for infection and disease in amphibians. *Diseases of Aquatic Organisms*, 101(3):235–242.

[Modha, 2003] Modha, D. S. (2003). Feature Weighting in k-Means Clustering . *Machine Learning*, 52(3):217–237.

[Moreau et al., 1999] Moreau, P., Anizon, F., Sancelme, M., Prudhomme, M., Bailly, C., Sevère, D., Riou, J.-F., Fabbro, D., Meyer, T., and Aubertin, A.-M. (1999). Syntheses and Biological Activities of Rebeccamycin Analogues. Introduction of a Halogenoacetyl Substituent. *Journal of Medicinal Chemistry*, 42(4):584–592.

[Muletz et al., 2012] Muletz, C. R., Myers, J. M., Domangue, R. J., Herrick, J. B., and Harris, R. N. (2012). Soil bioaugmentation with amphibian cutaneous bacteria protects amphibian hosts from infection by Batrachochytrium dendrobatidis. *Biological Conservation*, 152:119–126.

[Murtagh, 1983] Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms. *The Computer Journal*, 26(4):354–359.

[Ning et al., 2012] Ning, Y.-Y., Jin, D.-W., Sheng, G.-P., Harada, H., and Shi, X.-Y. (2012). Evaluation of the stability of hydrogen production and microbial diversity by anaerobic sludge with chloroform treatment. *Renewable Energy*, 38(1):253–257.

[North and Alford, 2008] North, S. and Alford, R. A. (2008). Infection intensity and sampling locality affect Batrachochytrium dendrobatidisdistribution among body regions on green-eyed tree frogs Litoria genimaculata. *Diseases of Aquatic Organisms*, 81:177–188.

[Orlov et al., 2008] Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D. M., and Goldberg, I. G. (2008). WND-CHARM: Multi-purpose image classification using compound image transforms .

[Oros-Peusquens et al., 2011] Oros-Peusquens, A. M., Matusch, A., Becker, J. S., and Shah, N. J. (2011). Automatic segmentation of tissue sections using the multielement information provided by LA-ICP-MS imaging and k-means cluster analysis. *International Journal of Mass Spectrometry*, 307(1-3):245–252.

[O'Toole et al., 2000] O'Toole, G., Kaplan, H. B., and Kolter, R. (2000). **Biofilm Formation As Microbial Developement** . *Annual Review of Microbiology*, 54(1):49–79.

[O'Toole and Kolter, 1998] O'Toole, G. A. and Kolter, R. (1998). Flagellar and twitching motility are necessary for Pseudomonas aeruginosa biofilm development. *Molecular Microbiology*, 30(2):295–304.

[Ouellet et al., 2005] Ouellet, M., Mikaelian, I., Pauli, B. D., Rodrigue, J., and Green, D. M. (2005). Historical Evidence of Widespread Chytrid Infection in North American Amphibian Populations. *Conservation Biology*, 19(5):1431–1440.

[Pamp et al., 2008] Pamp, S. J., Gjermansen, M., Johansen, H. K., and Tolker-Nielsen, T. (2008). Tolerance to the antimicrobial peptide colistin in Pseudomonas aeruginosa biofilms is linked to metabolically active cells, and depends on the pmr and mexAB-oprM genes. *Molecular Microbiology*, 68(1):223–240.

[Pantanella et al., 2006] Pantanella, F., Berlutti, F., Passariello, C., Sarli, S., Morea, C., and Schippa, S. (2006). Violacein and biofilm production in Janthinobacterium lividum. *Journal of Applied Microbiology*, 0(0):061120055200056–???

[Pelicic et al., 1997] Pelicic, V., Jackson, M., Reyrat, J. M., Jacobs, W. R., Gicquel, B., and Guilhot, C. (1997). Efficient allelic exchange and transposon mutagenesis in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10955–10960.

[Penatti et al., 2012] Penatti, O. A. B., Valle, E., and Torres, R. d. S. (2012). Comparative study of global color and texture descriptors for web image retrieval. *Journal of Visual Communication and Image Representation*, 23(2):359–380.

[Pessier et al., 1999] Pessier, A. P., Nichols, D. K., Longcore, J., and Fuller, M. S. (1999). Cutaneous Chytridiomycosis in Poison Dart Frogs (Dendrobates spp.) and White's Tree Frogs (Litoria Caerulea). *Journal of Veterinary Diagnostic Investigation*, 11:194–199.

[Phogot et al., 2001] Phogot, S. K., Gupta, R., Burma, P. K., Sen, K., and Pental, D. (2001). On the estimation of number of events required for saturation mutagenesis of large genomes . *Scientific Correspondence*.

[Pidot et al., 2013] Pidot, S. J., Coyne, S. e. b., Kloss, F., and Hertweck, C. (2013). Antibiotics from neglected bacterial sources. *International Journal of Medical Microbiology*.

[Pollock and Larkin, 2004] Pollock, D. D. and Larkin, J. C. (2004). Estimating the Degree of Saturation in Mutant Screens. *Genetics*, 168(1):489–502.

[Qi and Han, 2006] Qi, X. and Han, Y. (2006). Incorporating multiple SVMs for automatic image annotation. *Pattern Recognition*.

[Queiroz et al., 2012] Queiroz, K. C. S., Milani, R., Ruela-de Sousa, R. R., Fuhler, G. M., Justo, G. Z., Zambuzzi, W. F., Duran, N., Diks, S. H., Spek, C. A., Ferreira, C. V., and Peppelenbosch, M. P. (2012). Violacein induces death of resistant leukaemia cells via kinome reprogramming, endoplasmic reticulum stress and Golgi apparatus collapse. *PLoS ONE*, 7(10):e45362.

[Rollins-Smith et al., 2002] Rollins-Smith, L. A., Carey, C., Longcore, J., Doersam, J. K., Boutte, A., Bruzgal, J. E., and Conlon, J. M. (2002). Activity of antimicrobial skin peptides from ranid frogs against Batrachochytrium dendrobatidis, the chytrid fungus associated with global amphibian declines. *Developmental & Comparative Immunology*, 26(5):471–479.

[Ruiz et al., 2004] Ruiz, L., Domnguez, M. A., Ruiz, N., and Viñas, M. (2004). Relationship between clinical and environmental isolates of Pseudomonas aeruginosa in a hospital setting. *Archives of Medical Research*, 35(3):251–257.

[Sakuragi and Kolter, 2007] Sakuragi, Y. and Kolter, R. (2007). Quorum-Sensing Regulation of the Biofilm Matrix Genes (pel) of Pseudomonas aeruginosa. *Journal of Bacteriology*, 189(14):5383–5366.

[Shawe-Taylor and Sun, 2011] Shawe-Taylor, J. and Sun, S. (2011). A review of optimization methodologies in support vector machines. *Neurocomputing*, 74(17):3609–3618.

[Stapper, 2004] Stapper, A. P. (2004). Alginate production affects Pseudomonas aeruginosa biofilm development and architecture, but is not essential for biofilm formation. *Journal of Medical Microbiology*, 53(7):679–690.

[Stauff and Bassler, 2011] Stauff, D. L. and Bassler, B. L. (2011). Quorum Sensing in Chromobacterium violaceum: DNA Recognition and Gene Regulation by the CviR Receptor. *Journal of Bacteriology*, 193(15):3871–3878.

[Stehling et al., 2002] Stehling, R. O., Nascimento, M. A., and Falcao, A. X. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. In *the eleventh international conference*, pages 102–109, New York, New York, USA. ACM Press.

[Stewart et al., 2004] Stewart, P. E., Hoff, J., Fischer, E., Krum, J. G., and Rosa, P. A. (2004). Genome-Wide Transposon Mutagenesis of Borrelia burgdorferi for Identification of Phenotypic Mutants. *Applied and Environmental Microbiology*, 70(10):5973–5979.

[Symonds et al., 2008] Symonds, E. P., Trott, D. J., Bird, P. S., and Mills, P. (2008). Growth Characteristics and Enzyme Activity in . *Mycopathologia*, 166:143–147.

[Szeliski, 2010] Szeliski, R. (2010). *Computer Vision: Algorithms and Applications* . szeliski.org/Book/.

[Tamaoki et al., 1986] Tamaoki, T., Nomoto, H., Takahashi, I., Kato, Y., Morimoto, M., and Tomita, F. (1986). Staurosporine, a potent inhibitor of phospholipid/Ca++dependent protein kinase. *Biochemical and Biophysical Research Communications*, 135(2):397–402.

[Titilawo et al., 2015] Titilawo, Y., Obi, L., and Okoh, A. (2015). Antimicrobial resistance determinants of Escherichia coli isolates recovered from some rivers in Osun State, South-Western Nigeria: Implications for public health. *Science of The Total Environment*, 523:82–94.

[Tong et al., 2004] Tong, X., Campbell, J. W., Balázsi, G., Kay, K. A., Wanner, B. L., Gerdes, S. Y., and Oltvai, Z. N. (2004). Genome-scale identification of conditionally

essential genes in E. coli by DNA microarrays. *Biochemical and Biophysical Research Communications*, 322(1):347–354.

[Tremblay et al., 2007] Tremblay, J., Richardson, A.-P., Lépine, F., and Déziel, E. (2007). Self-produced extracellular stimuli modulate the Pseudomonas aeruginosa swarming motility behaviour. *Environmental Microbiology*, 9(10):2622–2630.

[Ursani et al., 2008] Ursani, A. A., Kpalma, K., and Ronsin, J. (2008). Texture features based on local Fourier histogram: self-compensation against rotation. *Journal of Electronic Imaging*, 17(3):030503–030503–3.

[Walbot, 2000] Walbot, V. (2000). Saturation mutagenesis using maize transposons. *Current Opinion in Plant Biology*, 3(2):103–107.

[Whitford and Schumacher, 1964] Whitford, L. A. and Schumacher, G. J. (1964). Effect of a Current on Respiration and Mineral Uptake In Spirogyra and Oedogonium. *Ecology*, 45(1):168.

[Woodhams et al., 2007] Woodhams, D. C., Ardipradja, K., Alford, R. A., Marantelli, G., Reinert, L. K., and Rollins-Smith, L. A. (2007). Resistance to chytridiomycosis varies among amphibian species and is correlated with skin peptide defenses. *Animal Conservation*, 10(4):409–417.

[Woodhams et al., 2011] Woodhams, D. C., Bosch, J., Briggs, C. J., Cashins, S., Davis, L. R., Lauer, A., Muths, E., Puschendorf, R., Schmidt, B. R., Sheafor, B., and Voyles, J. (2011). Mitigating amphibian disease: strategies tomaintain wild populations and controlchytridiomycosis. *Frontiers in Zoology*, 8(1):8.

[Xia et al., 2014] Xia, L., Zheng, X., Shao, H., Xin, J., and Peng, T. (2014). Influences of environmental factors on bacterial extracellular polymeric substances production in porous media. *Journal of Hydrology*, 519:3153–3162.

[Yada et al., 2008] Yada, S., Wang, Y., Zou, Y., Nagasaki, K., Hosokawa, K., Osaka, I., Arakawa, R., and Enomoto, K. (2008). Isolation and characterization of two groups of novel marine bacteria producing violacein. *Marine biotechnology (New York, N.Y.)*, 10(2):128–132.

[Yang et al., 2007] Yang, L. H., Xiong, H., Lee, O. O., Qi, S. H., and Qian, P. Y. (2007). Effect of agitation on violacein production in Pseudoalteromonas luteoviolacea isolated from a marine sponge. *Letters in Applied Microbiology*, 44(6):625–630.

[Zheng et al., 2014] Zheng, B., Yoon, S. W., and Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4):1476–1482.

[Ziebuhr et al., 1999] Ziebuhr, W., Krimmer, V., Rachid, S., Lossner, I., Gotz, F., and Hacker, J. (1999). A novel mechanism of phase variation of virulence in Staphylococcus epidermidis: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. *Molecular Microbiology*, 32(2):345–356.