# Genomic Analysis by Single Cell Flow Sorting

**Juno Choe**

**A dissertation submitted in partial fulfillment**
**of the requirements for the degree of**

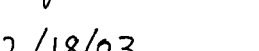
**Doctor of Philosophy**

**University of Washington**

**2003**

**Program Authorized to Offer Degree: Molecular Biotechnology**

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 12/18/03

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Juno Choe

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

Gerrit J. van den Engh

Reading Committee:

Gerrit J. van den Engh

Stanley Fields

Raymond J. Monnat, Jr.

Date: 12/10/03

University of Washington

**Abstract**

Genomic Analysis by Single Cell Flow Sorting

Juno Choe

Chair of the Supervisory Committee:
Affiliate Professor
Ger van den Engh
Department of Genome Sciences

The Human Genome Project has dramatically changed the landscape of biology. With the availability of genomic sequence from humans and many other organisms, new biological questions are being asked that involve the simultaneous study of thousands of genes or proteins. The invention of new technologies continues to be important for the timely investigation of many of these questions.

In this work, we present new technologies that address several genomics-level questions using electronic cell sorters. Because these machines are capable of examining and sorting tens of thousands of cells per second, they are potentially ideal platforms for investigating large systems. The challenge lies in converting biological attributes into readable physical attributes. In this work, we present the development of a series of plasmid vectors that encode biological states as the ratio of two fluorescent proteins in *E. coli*.

Using this doctrine, we created the pGRFP series of vectors that can be used to rapidly isolate insert-bearing clones on an electronic cell sorter. This technique is a powerful

alternative to traditional colony picking based on blue/white color selection. The speed of the electronic cell sorter allows us to deposit single cells into tubes as fast as the tubes can be transported. We validate this method's precision is selecting insert-bearing clones and show its usefulness in a small sequencing project.

We also show how the pGRFP series vectors can be used to classify a large number of protein mutants. We sequenced hundreds of active mutants of a human enzyme. From these data, we introduce the concept of the "x-factor" that indicates a particular protein's tolerance to mutation. We are able to make striking correlations between the pattern of mutability throughout the enzyme and what is known about its 3D structure and mechanism of action.

Finally, we present the pGFPpDsRed series of vectors that show promise in detecting DNA-Protein interactions. This might make a very useful tool for scanning genomic DNA for transcription factor binding sites on the road to solving regulatory networks. Conversely, a large number of protein mutants could be searched quickly to find variants that bind to a specific DNA sequence.

## TABLE OF CONTENTS

## LIST OF FIGURES

Figure Number Page

Figure Number Page

## LIST OF TABLES

Page

**Acknowledgements**

The author wishes to express his sincere gratitude to the following people who made this work possible.

Ger van den Engh has been a great mentor and friend throughout my graduate school experience. He has always supported and encouraged me to push my research into new and exciting directions.

My graduate supervisory committee consisting of Stan Fields, Ray Monnat, and Steve Moseley has always been eager to listen to new developments in my research. They have provided me with support and advice throughout my graduate school training. In addition, collaborations with members of the laboratories of Ray Monnat and Steve Moseley have been exciting and fruitful.

The department of Genome Sciences has provided a great environment for learning, research, and interaction with talented faculty and students.

I have really enjoyed working at the excellent facilities at the Institute of Systems Biology for the past several years. The research in systems biology occuring at the ISB has been a constant inspiration for my work.

I want to thank Haiwei Guo and Larry Loeb for the exciting and productive collaboration on the project described in the "Protein Mutagenesis Project" chapter. In addition to being a great friend to me over the years, Haiwei has proven to be a talented scientist to work with. Larry Loeb has always been supportive of our research and has provided countless examples of insightful comments that have furthered our research.

I want to thank Elijah Wallace in the Monnat Lab for our short but productive collaboration. Some of this research is described in the "Novel Method for Detecting DNA-Protein Interactions" chapter. That project would not have been possible without all of your hard work.

I am grateful to current and former members of the van den Engh lab including Alan Diercks, Tim Petersen, Monica Orellana, and Tom Schaus for making the lab such a great place to work. Their expertise in so many diverse fields has made this work possible. Their friendship has meant so much to me over the years.

Finally, I want to thank Sherrif Ibrahim for being there throughout all the trials and tribulations of the MD/PhD program. His sense of humor and perspective on life have helped keep me sane over the years. This work would not be possible without the countless discussions we had tossing around every sort of idea and experimental result.

# *Chapter 1. Introduction and General Background*

## *A) The Human Genome Project and the Age of Genomics*

Formally begun in 1990, the Human Genome Project was a planned fifteen year international collaboration to sequence the vast majority of the 3 billion bases comprising the human genome and to annotate the sequence in publicly accessible databases. To achieve these goals, the genome project went through several intermediary stages. The early stages of the project included mapping the human genome using sequence-tagged sites (STSs), short tandem repeat polymorphisms, and restriction fragment length polymorphisms. During the later stages of the genome project, the emphasis shifted to sequence tag connectors (STCs) from bacterial artificial chromosome (BAC) ends.

Throughout the Human Genome Project, the genomes of other organisms have also been sequenced along the way, leading to valuable experience in assembling the human genomic sequence. A robust microbial genome project has produced over 55 bacterial genomes to date, and the genomes of over 150 species are currently being sequenced. The multicellular organisms, *Caenorhabditis elegans* and *Drosophila melanogaster,* were sequenced as well.[1,2] As an adjunct to the human genomic sequence, a large number of expressed sequence tags (ESTs) from human cDNAs have also been sequenced to create expression databases.

---

[1] The C. elegans Sequencing Consortium. (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**: 2012-2018.
[2] Adams, M.D., S.E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**: 2185-2195.

A substantial investment was also made in technology development. Some of the innovations included better sequencing reagents, higher throughput capillary electrophoresis sequencing machines, and extensive robotics for colony picking and liquid handling in the pre-sequencing steps. In addition, many of the innovations were on the computational end, as new algorithms were required to compile enormous numbers of sequence reads into a contiguous sequence. New databases were needed to organize all of the sequence data and annotations into logical structures.

In 2001, the international public effort and Celera Genomics simultaneously published the rough draft sequence of the human genome.[3,4] This was a major milestone in the Human Genome Project, although some finishing work is still being done. Even with the completion of the Human Genome Project, the demand for sequencing throughput is expected to continue increasing. The construction of various human polymorphism databases will serve to correlate diseases and clinical treatment with individual gene variants. Ambitious sequencing projects for model organisms including chimpanzee, dog, rat, mouse, and fugu (pufferfish) have been initiated and are in various stages of completion. A strong sequencing program for microbiological organisms continues to announce completed genomes for significant bacteria, yeasts, and parasites. The Human Genome Project has not only sparked the need for sequencing throughput that was unimaginable decades ago, but has launched the

---

[3] The International Human Genome Mapping Consortium. (2001). "A Physical Map of the Human Genome." Nature **409**: 934-941.

[4] The Celera Genomics Sequencing Team. (2001). "The Sequence of the Human Genome." Science **291**: 1304-1351.

entire field of genomics that will continue to require high levels of DNA sequencing in the years to come.

As the Human Genome Project progresses, there is an evolving paradigm shift within certain sectors of the biological research community. Traditionally, biological questions are investigated one gene or protein at a time using conventional genetic and biochemical methods. With the prospect of having the entire human genomic sequence available as well as the sequence of all human genes through EST databases, it becomes clear that tools could be developed to investigate biology on entire genomic or proteomic scales. Some examples include microarrays for profiling the expression patterns of large numbers of genes under specific cellular conditions.[5] Advances in mass spectrometry and the development of isotope coded affinity tags (ICAT) has allowed the similar global comparison of protein levels.[6] The yeast two-hybrid system is amenable to high throughput detection of protein-protein interactions on a genome wide scale.[7] Besides these high profile examples, new schemes for more efficient methods of interrogating biology on a genomic scale are being developed continuously.

The invention of microarrays and the yeast two-hybrid system represent fundamentally new and faster ways of gaining the same information we gained previously using techniques such as Northern blots and co-immunoprecipitation.

---

[5] Lockhart, D.J., H. Dong et al. (1996). "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." Nat Biotechnol **14**: 1675-1680.

[6] Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**: 994-9.

[7] Fields, S. and O. Song. (1989). "A Novel Genetic System to Detect Protein-Protein Interactions." Nature **340**: 245-246.

There are other areas of biology that show incremental increases in efficiency brought about by extensive automation of established protocols. Instead of experiments done on several tubes at a time, 96-well or 384-well plates are used. Instead of hand pipetting reagents, robotic liquid handling systems process thousands of samples at a time. Instead of transferring PCR reactions between water baths, automated thermocyclers rapidly ramp the temperature of reactions. This automation is necessary to reduce human labor, making processes cheaper and more reliable. However, automation often reaches certain limits imposed by the chemistries or protocols themselves. Large increases in efficiency usually don't occur until new methods are invented that utilize different materials and methods.

In this work, we present a set of new methods and technologies for solving biological problems in the new genomic age. These biological methods are currently at the research stage and may not necessarily be faster or more reliable than commercially available technologies. However, our methods serve as a foundation that can be coupled with other biological insights and instrumentation to fuel the next generation of biological discoveries.

### *B) Clone Isolation by Flow Cytometry*

Although much biology can be analyzed by batch processing collections of cells, there is specific information that can only be attained by analyzing very narrow populations of cells or even single cells in some instances. It is often necessary to separate individual cells with unique genetic elements from within a complex

mixture. This could take the shape of trying to classify HIV infection among populations of fibroblasts.[8] B-cells from multiple myeloma patients were recently analyzed to isolate subpopulations of B-cells with a specific mutation in the N-Ras gene.[9] A complex mixture of naturally occurring bacteria from an environmental site was analyzed to discover its constituents.[10] Another application may be the haplotyping of individual sperm cells to look for trinucleotide repeat expansion in the Huntington's disease locus.[11]

Flow cytometers have been the tool of choice for decades to rapidly analyze and sort cells, one at a time. Briefly, a flow cytometer works by injecting cells of interest into the core of a thin jet of liquid approximately 70μm in diameter (see figure 1). Single or multiple lasers are typically focused onto this stream core where the cells pass. As cells fall past this point of focus, scattered light and/or fluorescence can be gathered by an array of carefully focused optics. Optical filters allow the user to select which emission wavelengths of fluorescence to observe. This light is detected by photomultiplier tubes which generate electrical pulses that correspond to each passing cell. Measured fluorescence values generally correspond to the number of fluorescent molecules are excited. Measured light scatter in the forward direction is a

---

[8] Bertram, S., F.T. Hufert, et al. (1995). "Detection of DNA in single cells using an automated cell deposition unit and PCR." Biotechniques **19**:616-620.
[9] Kalakonda, N., D.G. Rothwell DG, et al. (2001). "Detection of N-Ras codon 61 Mutations in Subpopulations of Tumor Cells in Multiple Myeloma at Presentation." Blood **98**:1555-1560.
[10] Bernard, L., C. Courties, et al. (2001). "A new approach to determine the genetic diversity of viable and active bacteria in aquatic ecosystems." Cytometry **43**:314-321.
[11] Chong, S.S., E. Almqvist, et al. (1997). "Contribution of DNA sequence and CAG size to mutation frequencies of intermediate alleles for Huntington disease: evidence from single sperm analyses." Human Molecular Genetics **6**: 301-309.

rough indication of the size of the cell. Light scatter in the 90 degree direction is indicative of the complexity of internal cellular structures.

Additional processes allow the isolation of desired cells from the sample. The flow cytometer nozzle is vibrated at a high frequency by a piezoelectric element, causing the jet stream to break into discrete droplets. When a cell of interest reaches the droplet breakoff point, a positive or negative charge is applied to the entire stream including the droplet containing that cell. When the droplet containing the cell actually breaks off, the entire stream is re-grounded to achieve electrical neutrality. The charged droplets are deflected by an electric field generated by two high voltage deflection plates and are collected in a receptacle below. Uncharged droplets continue to fall in a straight line and are collected in the waste chamber. Using this method, high speed cell sorters can reliably analyze cells at a rate of over 60,000 cells per second. They can currently sort desired cells at speeds up to 30,000 cells per second depending on purity requirements. However, if one desires to obtain many tubes containing only one cell each for clone isolation purposes, the speed with which tubes are transported in and out of the sort stream becomes the limiting factor. Some flow cytometers come equipped with methods of filling 96 well plates. However, this requires an operator to constantly replace each 96 well plate if a large number of samples must be processed. A convenient method of collecting and analyzing large numbers of single sorted cells still does not exist.

For this purpose, our laboratory is currently developing a linear format carrier tape for processing large numbers of samples. The current tape design contains 5000

wells of approximately 15 μl volume each. Reagents can be automatically deposited into each well along with sorted cells from a flow cytometer. A heat-sensitive cover tape can then be sealed over each well to isolate its contents (see figure 2a). This sealed tape can then be rolled up onto a reel, similar in action to a movie projector reel (see figure 2b). We have designed and built a tape handling instrument that can be interfaced to a flow cytometer (see figure 3). For each well, the instrument is capable of filling it with reagents, instructing the flow cytometer to deposit a given number of cells, and heat-sealing the well with the cover tape. We can currently fill 2-3 wells per second in this manner. At these speeds, we could isolate 7,000-10,000 clones per hour. If we implement this for both the left and right sort streams, speeds of 14,000 to 20,000 clones per hour could be achieved.

An adaptation of this tape handling instrument can transfer materials from one tape to another (see figure 4). This is useful for applications such as transferring cultured cells to PCR reactions for insert amplification. We have also developed a water-based thermocycler for use with reels of linear tape (see figure 5). The reaction chamber of the thermocycler is capable of holding 5 reels of tape (25,000 reactions). A pump sequentially transports water from one of three water baths into the reaction chamber. A computer monitors temperatures throughout the operation of the device and controls the series of pump and valves. This results in reliable thermocycling of all contents in the reaction chamber with extremely rapid temperature ramp rates.

This system is under development and serves as a clone isolation platform for the future. In one configuration, any given cell type could be identified by flow

cytometry and sorted by the thousands into wells on linear tapes. PCR reagents could be added to the wells, and sealed tapes could then be thermocycled. This would result in specific interrogations of genetic components on many thousands of separate cells within complex populations. To obtain even more detailed information, PCR products could undergo sequencing after being transferred to fresh tapes. Other configurations could be conceived to analyze flow cytometrically sorted clones in new and unique ways. In this proposal, a flow cytometry based bacterial clone isolation system is presented. This system can readily be used with traditional rectangular 96-well plates or the newly developed linear tape and handling instruments. Since the tape handling system is still under development, many of the experiments proposed here were performed initially in 96-well plates.

### C) *Bacterial Sorting for the Isolation of Recombinant DNA*

The ability to isolate one DNA species from a larger complex mixture is critical for all of molecular biology. This is a highly specialized form of clone isolation. Since single DNA molecules cannot be easily seen or handled, target DNA is usually ligated into a DNA vector that allows the entire construct to be propagated in some type of bacterial host strain. By isolating copies of host strain cells that originated from a single cell, separation of individual DNA species is achieved. The target DNA species can then be manipulated within the context of the larger construct. A large number of copies of the same DNA species can be isolated. Target DNA can be cut

out of a construct or alternatively, the target DNA can be copied out using PCR. A large number of options are available after clone isolation has been achieved.

Traditionally, *E. coli* bacteria have been the carriers of choice for this purpose. Usually, target DNA is ligated into a vector containing the lacZ-alpha gene. This vector is transformed into *E. coli*, and the cells are plated onto agar plates. The lacZ-alpha gene normally complements a mutated genomic copy of lacZ in the *E. coli* strain being used. By growing the bacteria on plates containing X-gal, functional beta galactosidase (lacZ) protein will convert X-gal into an insoluble blue dye; this results in blue color colonies. The presence of cloned inserts disrupts lacZ-alpha function leading to clear colonies. These clear colonies can then be picked using a sterile implement and transferred to liquid media or streaked out on another agar plate for growth.

This blue/white colony selection method has been used effectively for decades. In recent years, the use of automated colony picking robots has increased the speed with which blue/white colonies can be discriminated and picked (see figure 6A). There are several manufacturers of colony picking robots. The Genetix series of colony picking robots uses a camera to first image the plate. Then, the locations of all clear colonies are determined by software. A platform containing an array of metal pins is then transported across the agar plate. At each target, a metal pin pierces the colony, transferring the bacteria onto the pin (see figure 6B). This continues until all pins in the array have been exposed. The pins are then lowered into 96-well or 384-well plates containing culture medium. The array of pins is then automatically

washed in an ethanol bath and sterilized in a heating chamber before reuse. The whole process can then be repeated for the next set of 96 colonies. Using this method, the fastest Genetix model, the Megapix, can pick up to 4000 colonies per hour. Other colony picking robots are available from such companies as Genomic Solutions, Biorobotics, and Genemachines, but these solutions are generally slower.

Automated colony picking instruments have primarily been used by high-throughput sequencing operations. Modern high throughput sequencing facilities typically have dozens if not hundreds of sequencing machines. For instance, Celera currently uses more than 300 ABI PRISM 3700 capillary electrophoresis based sequencing machines. Each machine produces 96 sequencing reads approximately every two hours in largely automated operation. Therefore, at least 14400 sequence reactions must be prepared every hour to keep up with this enormous sequencing potential. The actual sequencing reactions have been heavily optimized and can be performed relatively easily with commercial reagents and thermocyclers after templates are prepared. The bottleneck has quickly become the process of isolating individual template clones from complex libraries as well as amplifying and purifying the DNA for sequencing. Advances in these areas could significantly improve sequencing throughput for a variety of different projects.

### *D) Summary of Contents*

This thesis describes the progress that has been achieved in creating novel technologies and tools for genomic analysis using flow sorting. The chapter entitled

"Sequencing Project" discusses unique vector technologies that use fluorescent proteins to allow cell sorters to isolate bacterial cells containing cloned inserts. These vectors are critical for converting biological events into physical properties that can be readily assayed with the proper instrumentation. We show the robustness of this vector technology by analyzing clones isolated by cell sorting from several randomized shotgun libraries. We further use our vectors to partially sequence a sample BAC from a shotgun library.

Although genomic shotgun sequencing and cDNA sequencing make up the largest demand for bacterial clone isolation, a number of other potential applications will require large numbers of bacterial clones in the future. Applications include various forms of protein engineering and mutagenesis. Several groups are dedicated to producing altered enzymes for industrial or clinical applications.[12] Generally, a library is produced containing thousands to millions of mutant variants of the wild type enzyme. Then, various forms of selection or screening are used to look for mutants with desirable qualities. At this point, a pooled library of functional mutants must undergo clone isolation, so that individual mutants can be analyzed. This further analysis could include additional functional screening or sequencing to determine the nature of the mutations induced. In the chapter entitled "Protein Structure Function Studies", we show how high speed bacterial sorting could significantly aid in the classification of large numbers of mutants. In collaboration with Haiwei Guo and Larry Loeb (University of Washington, Seattle, WA), we sequenced hundreds of

[12] Skandalis, A., L.P. Encell, and L.A. Loeb. (1997). "Creating Novel Enzymes by Applied Molecular Evolution." Chemistry & Biology **4**: 889-898.

active mutants of the human AAG enzyme. We invent the concept of a "x-factor" that indicates a particular protein's tolerance to mutation. We make striking correlations between the pattern of mutability throughout the AAG enzyme and what is known about its 3D structure and mechanism of action.

There are many other possible applications of high speed bacterial sorting that are only limited by the ingenuity of experimental design. In the most general sense, we can treat bacteria as the carriers for genetic material as well as protein factories that can create fluorescent signals that correlate somehow with the genetic material they carry. If some aspect of the genetic material cloned into a particular vector can be converted into fluorescence, then clones carrying desirable inserts can be isolated quickly by cell sorting. One example of this general thinking is the detection of DNA-Protein interactions. In the chapter entitled "Novel Method for Detecting DNA-Protein Interactions Background", we present a series of vectors that show promise in converting a DNA-Protein binding event into a specific fluorescence signal. By interacting a known DNA sequence with a library of mutant DNA binding proteins, we can find protein variants with a given binding specificity. Conversely, if we interact a library of DNA target sequences with a known protein, we can find binding sites for a given protein. One example of this would be the scanning of genomic DNA for sites where transcription factors bind.

The chapters entitled "Fluorescence Kinetics" and "RecA Independent Recombination" are theoretical chapters that discuss some of the pitfalls of the approaches explored in this thesis. "Fluorescence Kinetics" describes a model for

understanding the extreme variations in fluorescence of bacteria throughout various stages of culture growth. "RecA Independent Recombination" describes the potential for rearrangements within bacterial vectors, especially when there are large regions of similarity.

The final chapter is the conclusion for this thesis. In addition to concluding remarks, it describes an experiment involving the amplification of an entire genome from a single sorted cell. These types of studies have large implications for studying many un-culturable organisms present in our environment. This is the ultimate generalization of our concept of clone isolation and applies to all types of individual cell types. In this case, a clone is a specific type of organism that can be classified and examined based on its genomic contents.

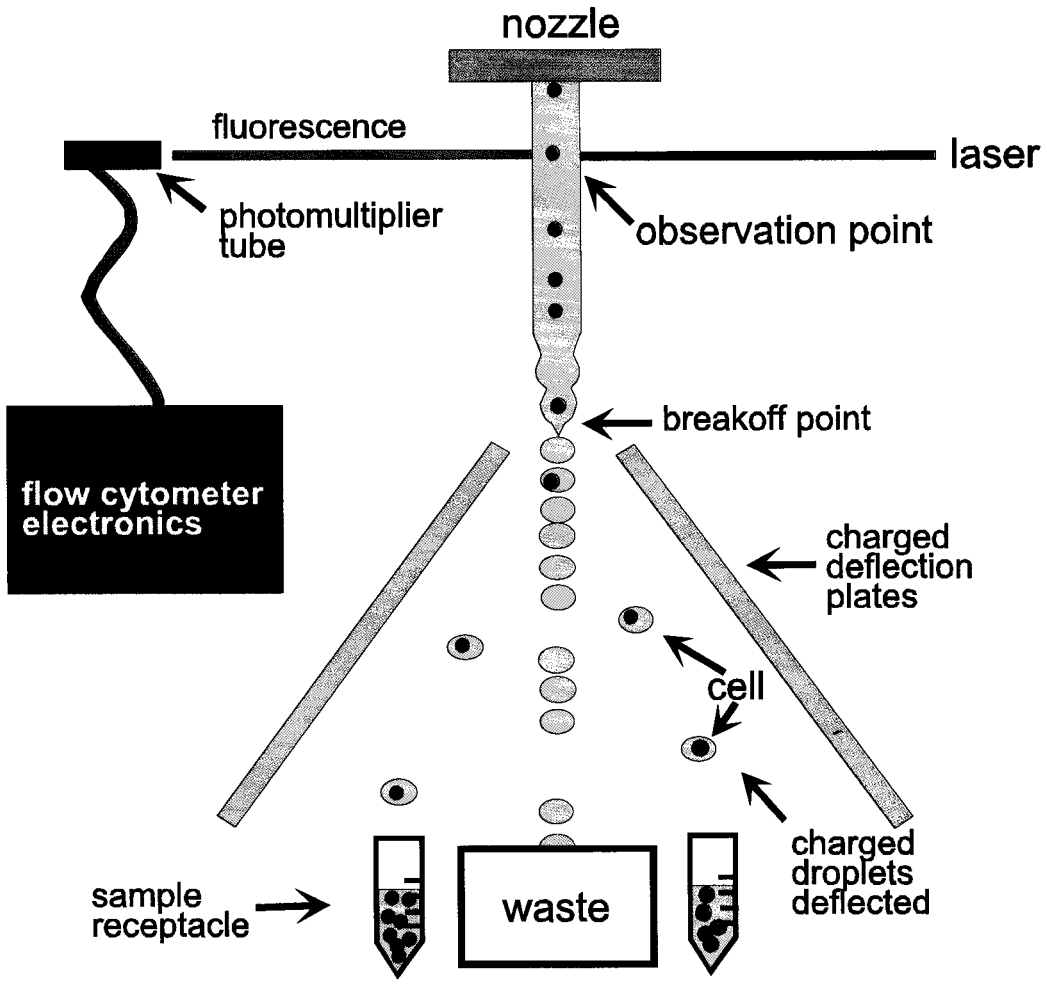
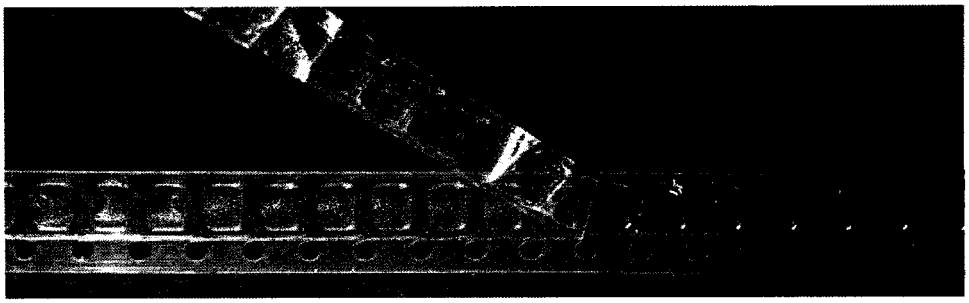



**Figure 1. General layout of a standard stream-in-air flow cytometer.** A laser hits the stream at the observation and interrogates scatter or fluorescence parameters for a cell. If the cell is desirable, a charge is applied to the entire stream just before the droplet containing the cell breaks off from the stream at the marked "breakoff point". The charged droplet is then deflected from the plane of the stream by two charged deflection plates, and the droplet is collected in a receptacle. The rest of the stream is re-grounded directly after the droplet breaks off. Uncharged droplets fall into a waste drain.

Adapted from web site: http://www.icb.ufmg.br/~prodap/projetos/cruzi/Facs.html.

A



B

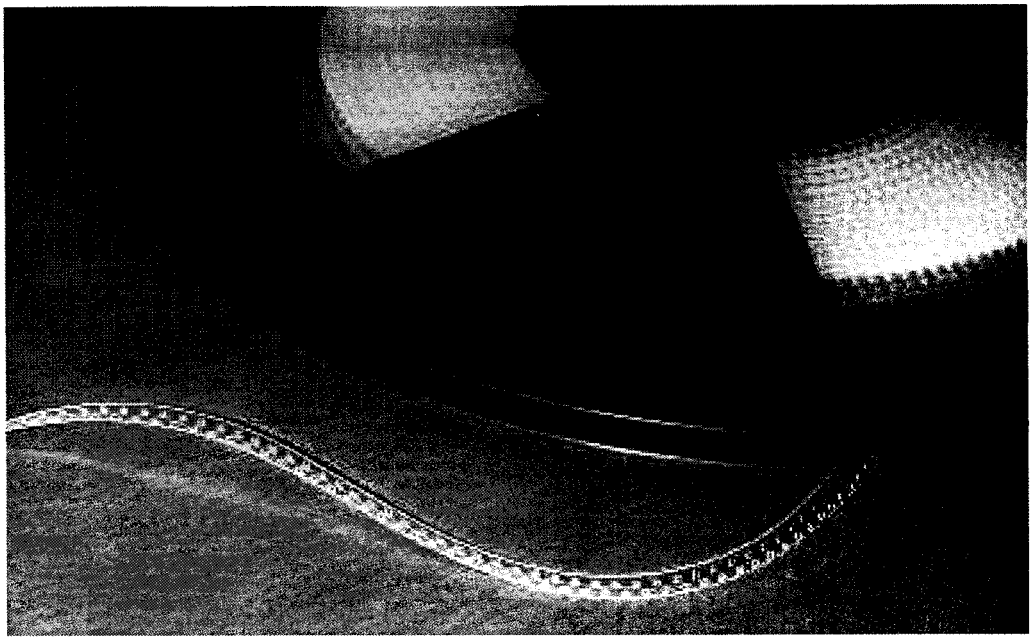
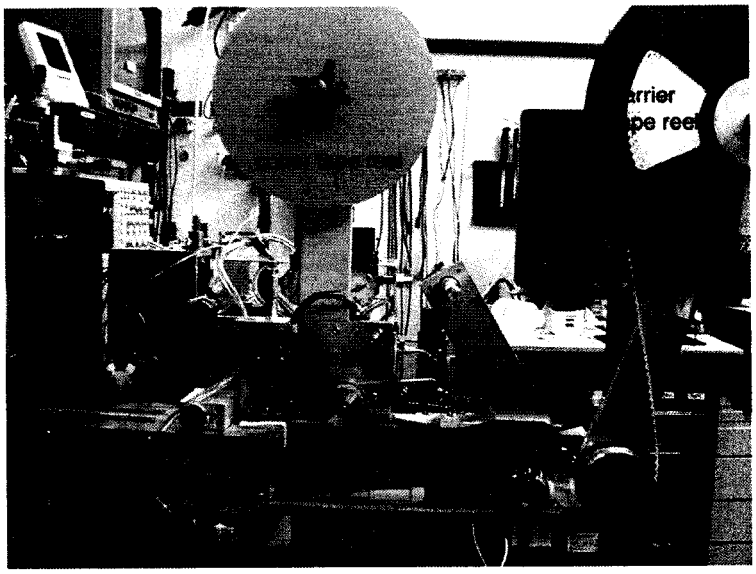


**Figure 2. Photograph of linear tapes.** A) A sample of linear tape is shown with approximate 15 µl wells. Heat-sensitive cover tape is shown after being sealed over several wells. The sprocket holes at the bottom allow instruments to precisely position the tape during handling. B) A larger length of linear tape is shown partially wound on a reel.

Photographs provided by Tom Schaus.

A



B

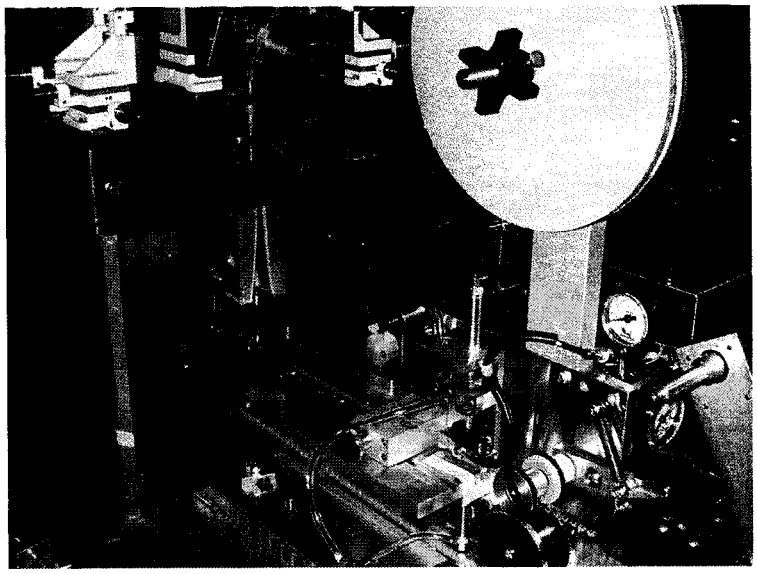


**Figure 3. Photographs of tape handling machine.** A) A side profile of the tape handling machine is shown attached to a flow cytometer. Empty tape is pulled from the carrier tape reel and transported along the bottom of the instrument and into the flow cytometer. A liquid pump deposits reagents into each well. Then, cells can be sorted into each well as it passes the position of the FACS sort stream. Finally, the tape wells are heat sealed to a cover tape at "sealer". B) More forward view shows details of coupling of the tape handling instrument to the flow cytometer.

Photographs provided by Tom Schaus.

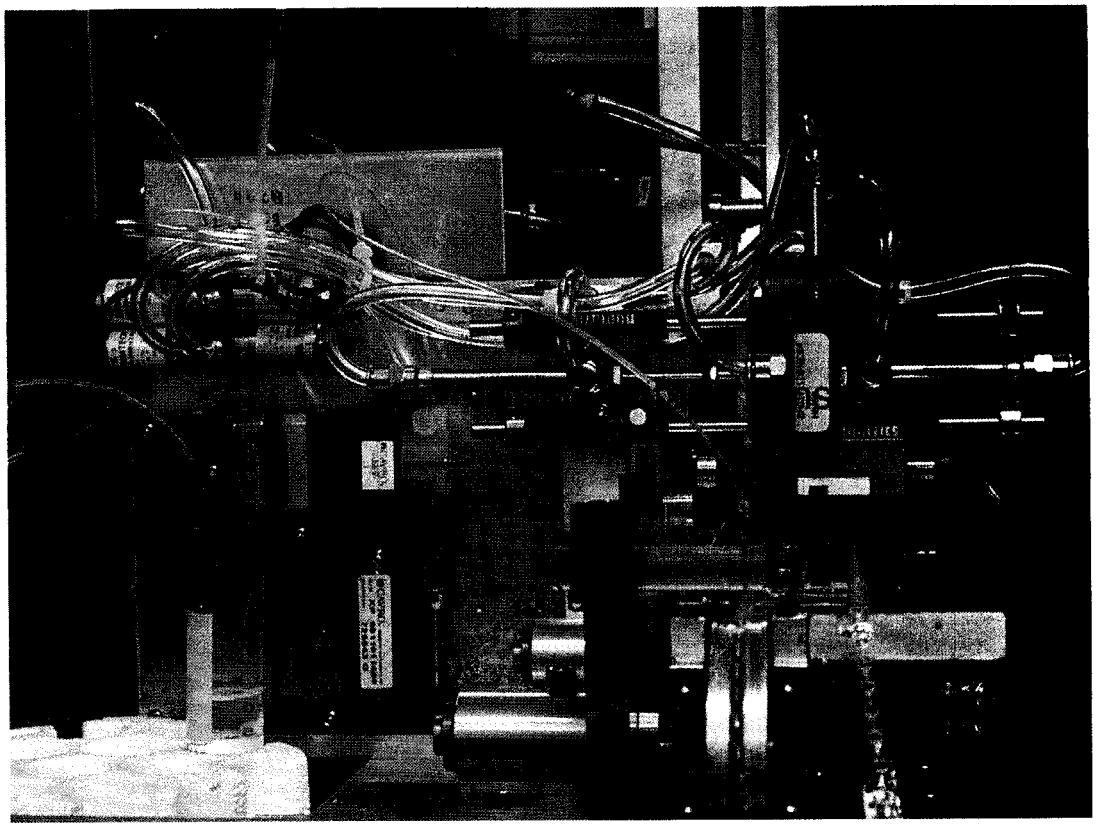


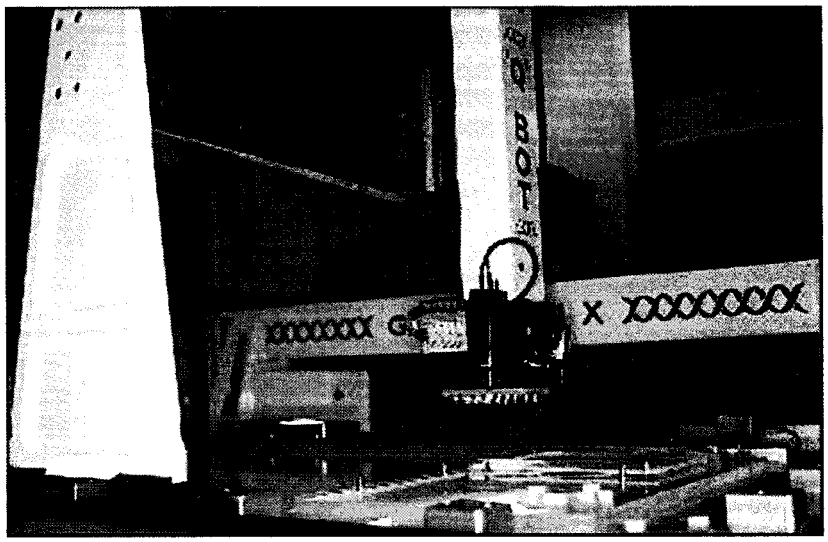
**Figure 4. Close-up photograph of tape handling instrument.** The tape handling instrument can be configured as a tape-to-tape transfer machine. Here, a pin marked "Probe" pierces wells from a culture tape and transfers small amounts of material to a PCR tape. A pump adds PCR reagents to each well at location marked "Outlet". The newly filled PCR tape is then sealed with heat-sensitive cover tape.

Photograph provided by Tom Schaus.


**Figure 5. Front view of custom built tape thermocycler.** The tapes to be thermocycled are placed in the reaction chamber (marked "PCR chamber"). A series of valves, shown underneath the reaction chamber, control the flow of water from three - 11 liter water baths (marked "Tanks") behind the primary water pump (marked "Pump"). The water enters the reaction chamber and rapidly alters the temperature of each tape well uniformly. Water is returned to the appropriate water bath after leaving the reaction chamber. The entire process is controlled by software running on a PC.

Photograph provided by Tom Schaus.

A



B

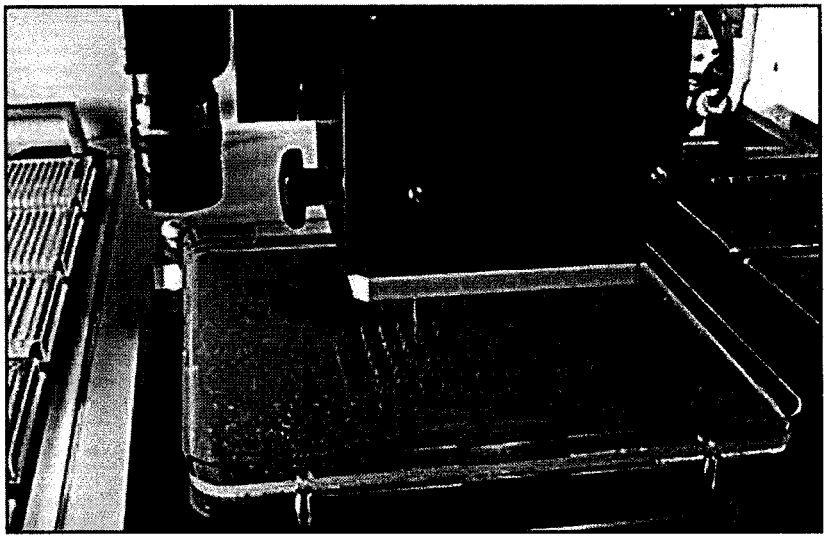


**Figure 6. Photographs of Genetix QPix robot.** A) The general layout of a Genetix QBot colony picking robot is shown. The head containing a camera and an array of pins can be transported between agar plates, 96-well plates containing culture medium, and a decontamination station. B) The head of a Genetix QPix robot is shown. The mounted camera for visualizing agar plates is shown on the left side of the head. The pin array is hovering over an agar plate, and a single pin can be seen piercing a colony.

Photographs reproduced from Genetix promotional web site: www.genetix.com .

# *Chapter 2. Sequencing Project*

## *A) Background – The Dual Fluorescent Protein Approach*

In this thesis, we address technologies that will allow us to utilize flow cytometers to more effectively isolate bacterial clones. Besides assays based on lacZ gene activity to indicate the presence of insert, several experimental systems have been developed based on fluorescence.[1] This is important to our discussion, because lacZ activity is not easily assayed under flow cytometry conditions, since flow cytometers generally observe only light scatter and fluorescence from cells. However, the groups developing these fluorescence based clone selection systems are utilizing their fluorescence in the context of colony picking on solid media. They cite the advantage that the fluorescence from each colony develops naturally without the addition of substrates such as X-gal used in lacZ assays. Inouye, et al. developed a plasmid vector that directs the production of GFP under a constitutive promoter.[2] When an insert is successfully ligated into the vector, stop codons within the insert prevent the successful translation of GFP. Therefore, non-fluorescent colonies could be picked to obtain individual clones. Roessner et al. developed a plasmid vector containing the uroporphyrinogen III methyltransferase (cobA) transcriptional marker

---

[1] Inouye, S., H. Ogawa, et al. (1997). "A Bacterial Cloning Vector using a Mutated Aequorea Green Fluorescent Protein as an Indicator." Gene **189**: 159-162.
[2] Ibid.

gene that produces a fluorescent product within *E. coli* cells.[3] Similar to the Inouye, et al. construct, colonies lose their fluorescence if an insert is present. These processes work well for colonies, but negative selection (the absence of fluorescence) is a poor criteria for the analysis of individual cells in a cell sorter. The degree of cell-to-cell variability in protein expression makes it impossible to guarantee that a weakly fluorescent cell is not expressing the fluorescent marker of interest. Additionally, background particles are largely weakly fluorescent, and it is difficult to segregate them from non-fluorescent bacteria. In order for sorting to be a viable approach, the creation of positive fluorescent markers is essential.

In this chapter, we will describe the construction of a series of novel cloning vectors that allow us to detect individual *E. coli* cells carrying inserts under flow cytometry conditions. We will describe some of the problems with the initial vector designs and how these problems were solved. The final vector designs indicate integration of a cloned insert through the shifting of ratios of two fluorescent proteins. In the example of the pGRFP vector, a fusion protein composed of Green Fluorescent Protein and DsRed Fluorescent Protein is produced. The GFP/DsRed fusion protein can be excited by blue light at 488nm. This causes excitation of the GFP, followed by DsRed due partly to direct excitation and partly to fluorescence resonance energy transfer (FRET). When an insert is successfully integrated into the cloning site between GFP and DsRed, there is a loss of function of the DsRed portion of the

[3] Roessner, C.A. and A. Ian Scott. (1995). "Fluorescnence-Based Method for Selection of Recombinant Plasmids." Biotechniques **19**: 760-764.

protein. In nearly all cases, the DsRed protein is not translated due to the insertion of an up-stream stop codon. Increased green fluorescence will be observed with loss of observable red fluorescence. We can quantitate the ratios of these two fluorescent proteins very accurately by flow cytometry, leading to reliable separation of individual GFP+/DsRed- bacteria from the total pool of transformation products. The system is immune from the large variance in expression levels, since we are basing our sort decisions only on the ratio between the two fluorescences. Testing indicates that this system is a robust and rapid method of isolating insert-containing bacterial clones from a liquid culture.

### *B) Vector Development and Testing – pSE-GFP series vectors*

Our goal is to design vectors that allow *E. coli* cells to be differentiated by flow cytometry depending on the presence or absence of insert. Flow cytometers are capable of quantitating fluorescence and forward light scatter levels from bacteria as elicited by a laser light source. However, it is relatively difficult to alter forward scatter levels of bacteria, because this parameter is directly linked to cell size and orientation of the rod shaped cells. Fluorescence is a more tractable parameter that can be coupled to genetic changes. For this purpose, fluorescent proteins are an obvious choice, because they are directly coded by genetic elements and can be produced endogenously within *E. coli* without the need for added substrates. The first vector designs involved single fluorescent proteins. Vectors could be designed to either upregulate or downregulate fluorescence in response to the presence of an

insert. Knocking out fluorescence in the presence of an insert is relatively straight forward. The cloning site for the insert can reside either in the promoter or coding region of the fluorescent protein. The insert will either break up the promoter so the gene is not transcribed, or the insert will cause premature termination of translation if cloned into the middle of the fluorescent protein coding sequence. However, it is more reliable to flow cytometrically sort cells that have upregulated fluorescence in response to an insert. This prevents the inadvertent sorting of dead cells or other contaminating non-fluorescent cell types.

Our first vector design tried to implement this upregulation of fluorescence in the presence of an insert. The final evolution of this design was the 4880bp pSE-GFP+ vector (see figure 7A). A bright mutant of GFP (GFPmut3.1, Clontech Laboratories, Inc.) was cloned downstream of a strong modified A3 promoter[4] and a symL20 symmetric high affinity lac operator site.[5] The modified A3 promoter provides better regulation of transcription; the symL20 lac operator site has many fold higher affinity for lacIq than the wildtype operator site. The lac repressor (lacIq) gene was also expressed at high levels from the same plasmid using a constitutive A3 promoter. Within this promoter, we included a cloning site for insert DNA as well as flanking M13 forward and reverse priming sites (see figure 7B). A rrnB T2 transcriptional terminator was also placed between the lac repressor gene and the GFP

[4] Yamada, M., M. Kubo, et al. (1991). "Promoter Sequence Analysis in Bacillus and Escherichia: Construction of Strong Promoters in E. coli." Gene **99**: 109-114.
[5] Sasmor, H. and J. Betz. (1990). "Symmetric lac operator derivatives: Effects of Half-operator Sequence and Spacing on Repressor Affinity." Gene **89**:1-6.

gene to prevent read-through transcription from the lacIq promoter into GFP. There are several advantages of cloning foreign DNA inside the promoter region of a gene compared with inside a coding region. There is less sensitivity to small one or two base deletions that can be caused by stray exonucleases during the cloning process. Also, since the insert is never transcribed or translated, we predict less clone bias will exist due to the presence of different inserts.

In the native pSE-GFP+ plasmid, the production of lacIq keeps GFP expression levels at a minimum. In plasmids with inserts, the –10 and –35 regions of the lacIq gene are separated so that they are no longer recognized by the RNA polymerase. The lacIq gene is no longer produced, so GFP protein is produced unhindered. This leads to increased fluorescence in *E. coli* containing inserts compared with non-insert containing bacteria. Early testing of the vector yielded promising results. A definite difference in brightness could be observed by fluorescence microscopy between cultures of *E. coli* containing pSE-GFP+ with and without an insert (see figure 8). By flow cytometry, a difference could be seen between insert containing and non-insert containing *E. coli* in culture (see figure 9A&B). We also analyzed a culture from bacteria transformed with a library of inserts cloned into pSE-GFP+. This culture consisted of both insert containing and non-insert containing bacteria (see figure 9C). However, a definitive boundary could not be established that separated these two populations. Each population has so much possible variance that the overlap between the two is too great.

This setback highlights the extreme difficulty in controlling expression levels within individual *E. coli* cells. Under microscopy, *E. coli* often exhibit large variations in length. Some of the rod-shaped bacteria appear long compared with other cells. Some of these longer cells may be in the process of replication. In any case, this can affect the total amount of fluorescence as observed by flow cytometry. In addition to size, there are also large variations in fluorescent protein concentrations from cell to cell. Elowitz et al. studied the kinds of noise that exist within bacterial cells expressing two fluorescent proteins.[6] They describe two kinds of noise, "intrinsic" and "extrinsic" that lead to fluctuations in protein expression. "Intrinsic noise" is responsible for variations in the expression of different proteins within a given cell. "Extrinsic noise" is responsible for global differences in overall protein expression from cell to cell.

In many cases, a culture grown from a single colony exhibits very heterogeneous characteristics. This can be seen in figure 9B in which the culture was grown from a single pSE-GFP+ insert containing colony. Numerous cells can be seen in the image with a wide range of fluorescence intensities. Further illustration of this point is provided by flow cytometry analyses of *E. coli* cultures containing pGFPmut3.1, a plasmid that expresses GFP from a constitutive promoter (see figure 10). At stationary phase, a huge variation in expression levels is observed.

---

[6] Elowitz, M.B., A.J. Levine, E.D. Siggia, and P.S. Swain. (2002). "Stochastic Gene Expression in a Single Cell." Science **297**: 1183-1186.

### *C) Vector Development and Testing – pBGFP series vectors*

To address the problem of variability in expression levels, we needed a fluorescent protein scheme that did not rely on expression levels at all. We created pBGFP, a novel 3809bp sequencing plasmid. (see figure 11). This vector indicates integration of a cloned insert through the shifting of ratios of two fluorescent proteins. In the native pBGFP, a fusion protein composed of Blue Fluorescent Protein (BFP2, Clontech Laboratories, Inc.) and GFPmut3.1 is expressed from a lac promoter, and translation is enhanced by a consensus ribosomal binding site. The BFP/GFP fusion protein can be excited by UV light in the 350-400nm wavelength range; this causes excitation of both the BFP and GFP fluorophores. We also believe that some GFP fluorescence is due to partial fluorescence resonance energy transfer (FRET) between the two fluorophores. When an insert is successfully ligated into the cloning site in the linker region between BFP and GFP, the introduction of stop codons terminates the translation of the protein prior to reaching the GFP half of the protein. In this case, we expect to see the loss of observable GFP fluorescence. At the same time, we expect increased BFP fluorescence due to the disruption of FRET between BFP and its acceptor partner GFP. Therefore, a large difference in the BFP:GFP fluorescence ratio exists between bacteria containing insert compared with bacteria without insert. Because we can quantitate this fluorescence ratio very accurately by flow cytometry, we have the ability to rapidly differentiate and sort only individual bacteria containing an insert. It should be noted that no reliance is made on expression levels, since we

base sort decisions only on the ratio of the two fluorescences. As long as the expression levels are high enough to accurately quantitate the relative numbers of the two types of fluorophores, an accurate sort decision can be made.

The pBGFP vector is diagramed in figure 11. Most of the vector including the BFP portion of the fusion protein was derived from pBFP2 (Clontech Laboratories, Inc.). We cut pBFP2 with BsmI to destroy the BFP2 gene stop codon. The linear fragment was treated with Mung Bean Nuclease to create a blunt end. This fragment was then digested with SpeI to a sticky end on the other side. Both ends were 5' dephosphorylated with Calf Alkaline Intestinal Phosphatase (CIP).

The GFPmut3.1 gene was PCR amplified from pGFPmut3.1 (Clontech Laboratories, Inc.) using one mutagenic primer at the upstream end in order to generate the linker region. This mutagenic primer included the linker region sequence as well as the first 26 bases of the GFPmut3.1 gene. The opposing primer was a standard oligonucleotide that primed downstream of the GFPmut3.1 gene. This PCR product was purified and then digested with SpeI to generate one sticky end.

The two fragments were ligated using T4 DNA Ligase to generate a prototype pBGFP vector. The ligation products were electrotransformed into Electromax DH10B cells (Gibco BRL) and plated onto ampicillin plates. Colonies expressing GFP were picked, and the clones were verified by sequencing.

In order to increase expression levels, we added a consensus *E. coli* Shine-Dalgarno (SD) sequence[7,8,9] with the proper 7bp spacing between the end of the SD sequence and the Blue Fluorescent Protein start codon. In the process, we also deleted a spurious M13 reverse priming site present on the original pBFP2 plasmid. We started by cutting the prototype pBGFP with Pci I and Age I. We used Mung Bean Nuclease to create blunt ends on both sides. We then dephosphorylated the two ends with CIP and gel purified the fragment. We PCR amplified a 125 bp fragment from the promoter region of the prototype pBGFP upstream of the blue-green fusion protein. We used the mutagenic primer containing the new SD sequence in the reverse direction. The PCR product was purified and ligated with the blunt-ended Pci I, Age I fragment of the prototype pBGFP. The ligation products were again electrotransformed into Electromax DH10B cells and plated onto ampicillin plates. Colonies expressing fluorescent protein were picked, and the clones were verified by sequencing the promoter region of the blue-green fusion protein. For some experiments, a later version of the pBGFP vector was used containing a Not I and Sal I sites in addition to the EcoRV site in the linker region between the BFP and GFP genes.

---

[7] Chen, H., M. Bjerknes, R. Kumar, and E. Jay. 1994. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nucleic Acids Research **22**: 4953-4957.

[8] Vellanoweth, R.L. and J.C. Rabinowitz. 1992. The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. Molecular Microbiology **6**: 1105-1114.

[9] Tang, G.L., Y.F. Wang, J.S. Bao, and H.B. Chen. 1999. Overexpression in Escherichia coli and Characterization of the Chloroplast Triosephosphate Isomerase from Spinach. Protein Expression and Purification **16**:432-439.

To create a random shotgun library of Adenovirus-2 DNA, commercially available purified Adenovirus-2 DNA was subjected to sonication in 1X Mung Bean Nuclease Buffer. End repair was performed with Mung Bean Nuclease for ~45 minutes at 30 degrees. Fragments with lengths 2-3kb were isolated by gel purification. The pBGFP plasmid was cut with EcoRV and subjected to dephosphorylation by CIP. The linearized vector fragment was isolated by gel purification. The purified Adenovirus-2 fragments were blunt-end ligated to linearized pBGFP vector. Ligation products were electrotransformed into Electromax DH10B cells and incubated in 1ml SOC medium for 1 hour according to standard protocols. We inoculated 5 ml. fresh LB medium plus 100 μg/ml ampicillin with 100 microliters of the SOC mix containing transformants. After ~17 hours of growth at 37oC, 225RPM rotary shaking, cells were harvested by centrifugation. The pellet was resuspended in 0.9% NaCl for analysis by flow cytometry or fluorescence microscopy.

Fluorescence microscope Images were obtained by observing 10μl of culture resuspended in 0.9% NaCl on a glass slide using a Zeiss Axiophot fluorescence microscope. Illumination was with arc lamp light filtered to the long-wave UV range. Images were captured using the standard Zeiss 35mm camera with 850 ISO color slide film and 1-3 second exposure times. Slides were developed and scanned using an AGFA slide scanner. The fluorescence microscopy images of *E. coli* grown with DNA libraries cloned into pBGFP highlight the relatively large difference in

fluorescence profiles between insert containing and non-insert containing bacteria (see figure 12). The bacteria containing native pBGFP appear green due to the relative brightness of the GFP fluorophore and the partial FRET between the blue and green fluorophores. The insert containing bacteria appear blue due to the loss of translation of the GFP fluorophore.

Flow cytometry was performed using an Influx flow cytometer (Cytopeia, Inc.). The sample cells were analyzed for blue and green fluorescence according to the setup shown in figure 3. Excitation was by a 5W Innova argon laser (Coherent) set to multi-line UV emission at ~250mW power. Fluorescence was passed through a 425 long pass filter to remove artifacts from the excitation light. The fluorescence was then passed through a 490 dichroic long pass beamsplitter to separate blue and green fluorescence signals. The blue fluorescence was further filtered through a 440/80 band pass filter before being detected by a photomultiplier tube. The green fluorescence was passed through a 520 long pass filter before its respective photomultiplier tube. The flow cytometer was set to trigger off of blue fluorescence. The Green Fluorescent Protein signal was plotted vs. the intensity of the Blue Fluorescent Protein signal on a linear bivariate plot for each analyzed sample (see figure 13). A clear difference in ratios between two populations can be seen. In the fused BFP/GFP protein, an exact 1:1 ratio of the two fluorophores is present due to the fused nature of the protein. With an insert, only the blue fluorophore is present. In

both of these cases, extremely tight distributions of fluorescence ratios are measured, leaving a large separation between the two populations.

Cells that showed high blue fluorescence versus green fluorescence were selected for sorting. The flow cytometer was programmed to deposit a single event into each well of a 96 well plate containing 200 microliters LB (plus 50 micrograms/ml. ampicillin) per well. These plates were grown overnight at 37 degrees. Typically, approximately 60% of the wells had growth that was visibly apparent. Sterile pipette tips were dipped into 20 cultures with growth and then used to inoculate PCR reactions prepared with Biolase polymerase (Bioline USA, Inc.) using standard protocols with 2mM $MgCl_2$. Primers were designed to flank the linker region of pBGFP. The PCR products were run on a 1% agarose gel (see figure 14). Out of 20 clones, 13 clones had a strong PCR product representing amplification of the insert. We further analyzed the seven clones without PCR product and determined that 3 clones probably had inserts that were not amplified, 2 clones had deletions consistent with exonuclease activity during cloning, and 2 clones had major deletions consistent with recombination. These data are presented and discussed in a later chapter entitled "RecA Independent Recombination". Because of this high frequency of recombination even in RecA minus strains, we were forced to abandon the pBGFP vector design.

### *D) Vector Development and Testing – pGRFP series vectors*

To deal with the high frequency of recombination in pBGFP, we created a 3441bp plasmid vector, pGRFP, that incorporates a fusion protein comprised of GFPmut3.1 and wtDsRed[10] (Discosoma Red Fluorescent Protein). pGRFP is diagramed in figure 15. In pGRFP2, we substituted wtDsRed with DsRed-T3[11], which has faster fluorescence maturation characteristics. When DsRed was first isolated, it was noted that there was less than 30% amino acid homology between DsRed and GFP.[12] Comparing the nucleotide sequences of the two genes, no significant stretches of similarity could be found. Therefore, we expected a greatly reduced occurrence of rearrangements.

To construct pGRFP, the pUC vector backbone from a standard pUC plasmid was PCR amplified. This 1990 bp PCR product contained the origin, an AmpR marker, and a lac promoter. A consensus *E. coli* Shine-Dalgarno (SD) sequence[13,14,15]

---

[10] Matz, M.V., A.F. Fradkov, et al. (1999). " Fluorescent Proteins from Nonbioluminescent Anthozoa species." Nature Biotech **17**:969-973.

[11] Bevis, B.J. and B.S. Glick. 2002. Rapidly maturing variants of the Discosoma red fluorescent protein (DsRed). Nature Biotechnology **20**:83-87.

[12] Matz, M.V., A.F. Fradkov, et al. (1999). " Fluorescent Proteins from Nonbioluminescent Anthozoa species." Nature Biotech **17**:969-973.

[13] Chen, H., M. Bjerknes, R. Kumar, and E. Jay. 1994. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nucleic Acids Research **22**: 4953-4957.

[14] Vellanoweth, R.L. and J.C. Rabinowitz. 1992. The influence of ribosome-binding-site elements on translational efficiency in Bacillus subtilis and Escherichia coli in vivo. Molecular Microbiology **6**: 1105-1114.

[15] Tang, G.L., Y.F. Wang, J.S. Bao, and H.B. Chen. 1999. Overexpression in Escherichia coli and Characterization of the Chloroplast Triosephosphate Isomerase from Spinach. Protein Expression and Purification **16**:432-439.

and a 7 bp spacer were also engineered into the vector downstream from the lac promoter using a 5' tail on one of the primers. The PCR product was cleaned, cut with AatII, and treated with CIP. This yielded one blunt end and one AatII end, both 5' dephosphorylated.

The GFPmut3.1 gene was PCR amplified from pGFPmut3.1 (Clontech Laboratories, Inc.) starting at the start codon and ending just prior to the stop codon with an additional 56bp tail containing a 6 amino acid linker sequence, M13 forward priming site, a stop codon, a BsmI site and an AatII site. The forward primer at the GFP start codon was phosphorylated prior to PCR. The 868 bp PCR product was cleaned and cut with AatII.

Both DNA fragments were gel purified and ligated together with T4 DNA Ligase to generate a precursor vector. The ligation products were electrotransformed into Electromax DH10B cells (Invitrogen) and plated onto ampicillin plates. Colonies expressing GFP were picked, grown, miniprepped, and sequenced in the lac promoter and GFP stop codon region. A clone with proper sequence was cut with BsmI to destroy the GFP stop codon and treated with Mung Bean Nuclease and CIP to blunt and dephosphorylate the ends.

To construct the second portion of the vector, the DsRed gene from pDsRed (Clontech Laboratories, Inc.) was PCR amplified. The upstream primer contained a 5' end incorporating a NotI site, EcoRV site, SalI site, an M13 reverse priming site and another 6 amino acid linker sequence. The downstream primer contained a StuI site after the DsRed stop codon. The PCR product was cleaned and treated with DpnI to

digest template DNA. The product was further gel purified and blunt ligated to the GFP containing vector described above. After transformation, colonies showing both green and red fluorescence after 36 hours of growth were picked. These colonies were grown, miniprepped, and sequenced in the region of the linker between GFP and DsRed. One of the clones bearing the correct sequence was labeled pGRFP for downstream experiments.

To generate pGRFP2, the wtDsRed gene was replaced with a new DsRed-T3 mutant obtained from Glick et al. Most of the pGRFP vector was PCR-amplified including the GFP gene, linker, vector backbone, and several bases at the end of wtDsRed. The DsRed-T3 gene was PCR amplified from a plasmid containing this gene. Primers were pre-phosphorylated with T4 Polynucleotide Kinase. The PCR products were gel purified and ligated together. The ligation products were transformed, and the cells were plated and grown at 37 degrees. Colonies that developed visible red fluorescence after 24 hours were picked. One of the clones bearing the correct sequence was labeled pGRFP2 for downstream experiments.

Shotgun libraries were purified from commercially prepared Adenovirus-2 DNA (Gibco BRL), the murine BAC MB01F09, and the sea urchin BAC SU66E20. The BACs were prepared using either a modified alkaline lysis, phenol/chloroform extraction and digestion with Plasmid-Safe ATP-Dependent DNase (Epicentre) or by using a Qiagen Large Construct Plasmid Isolation Kit. The template DNA was subjected to sonication, and end repair was performed with either Mung Bean

Nuclease or treatment with T4 DNA Polymerase, Klenow Fragment, and T4 Polynucleotide Kinase. Fragments of approximately 1.5-3kb length were isolated by gel purification.

The pGRFP or pGRFP2 plasmids were cut with EcoRV enzyme and subjected to dephosphorylation by CIP. The linearized vector fragment was gel purified. The purified library inserts were blunt-end ligated to linearized pGRFP or pGRFP2. Ligation products were electrotransformed into Electromax DH10B cells and incubated in 1ml SOC medium for 1 hour according to standard protocols. 5 ml. fresh LB medium plus 50μg/ml carbenicillin was inoculated with 100 microliters of the SOC mix containing transformants. Cells were grown at 30°C, 225 RPM rotary shaking for 36 hours for pGRFP and 24 hours for pGRFP2.

A culture transformed with pGRFP and shotgun library inserts from the MB01F09 BAC was harvested for cells that were fixed on a cover slip with 10% formaldehyde. A Zeiss Axiophot fluorescence microscope with an attached Nikon Coolpix digital camera was used to photograph the cells. Illumination was with arc lamp light filtered to the blue range (450-490nm). Color digital images were transferred to a computer, and brightness and contrast were enhanced. Figure 16 shows one image from this series. The cells with native vector are orange, because they express both the green and red fluorescent proteins. The insert containing cells can be recognized by their green color.

Flow cytometry was performed using an Influx flow cytometer (Cytopeia, Inc.). Approximatey 10-50 µl of culture was added to 4 ml. of 0.9% NaCl, injected into a flow stream comprised of autoclave sterilized 0.9% NaCl. The sample cells were analyzed for green and red fluorescence according to the setup shown in Figure 3. Excitation was by a 5 W Innova argon laser (Coherent) set to 488nm emission with ~400mW power. The flow cytometer was configured to trigger either on forward scatter or on the green fluorescence signal. Fluorescence passed through a 488 rejection band filter and then was split using a 550 dichroic long pass beam splitter. The green fluorescence was filtered through a 560 short pass filter before detection. The red fluorescence passed through a 590 long pass filter before detection. Figure 17 shows bivariate dot plots of DsRed fluorescence versus GFP fluorescence for analyzed *E. coli* cells. The native pGRFP vector yields cells with a high DsRed to GFP ratio. When Adenovirus-2 fragments are cloned into pGRFP, a second population of cells appears that has a lower DsRed to GFP ratio. A large separation is observed between the two populations according to the ratio of DsRed to GFP fluorescence allowing a high degree of discrimination between the two populations.

Figure 18 shows the gradual maturation of red fluorescence with the pGRFP vector over a 36 hour time course. After 14 hours, the fused GFP/DsRed protein primarily exhibits green fluorescence with very little fluorescence contribution from DsRed. Slowly, over 24 more hours, DsRed fluorescence increases. This highlights the greatly increased maturation time required for DsRed as compared with GFP. The

DsRed-T3 mutant obtained from Bevis and Glick has a much faster fluorescence maturation time. This is relevant to our later discussion on the study of fluorescence kinetics.

Cells from the population with low red fluorescence versus green fluorescence were selected for sorting. The flow cytometer was programmed to deposit a single cell into each well of a 96-well plate containing 200 ml LB (plus 50 μg/ml carbenicillin and 0.5% glucose) per well. These plates were grown with single cells sorted from a culture containing pGRFP with cloned 1.5-3kb Adenovirus inserts. Typically, 50-70% of the wells had growth that was visibly apparent. Sterile pipette tips were dipped into 48 cultures showing growth and then used to transfer a small number of cells to PCR reactions prepared with Herculase DNA Polymerase (Stratagene) using manufacturer recommended protocols. Forward and reverse primers flanking the linker region of pGRFP were used. Figure 19 shows the PCR products as analyzed on a 0.8% agarose gel. Note that almost every PCR product has a single product with a unique length. This demonstrates the concept of clone isolation by single cell flow sorting. Of the 48 cultures, 46 PCR reactions yielded a PCR product consistent with a cloned insert. 44 clones had inserts in the 1.5-3kb size range of the original Adenovirus-2 library; the other two clones had inserts that were slightly smaller than 1kb. One PCR reaction yielded no PCR product and the other had an inconclusive PCR product.

We proceeded with the sequencing of a sample BAC, SU66E20, containing subcloned sea urchin DNA. The entire process is summarized in figure 20. A size selected shotgun library from SU66E20 in the 2.5-5kb size range was created as described above. The BAC was purified initially by alkaline lysis, phenol/chloroform extraction, and digested with Plasmid-Safe ATP-Dependent Dnase (Epicentre). The BAC DNA was sonicated and end repaired with T4 DNA Polymerase (Fermentas), Klenow Fragment (Fermentas), and T4 Polynucleotide Kinase (Fermentas). Fragments of approximately 2.5-3kb length were isolated by gel purification. The library inserts were blunt-end cloned into pGRFP2 and grown in culture at 30 degrees C for 28 hours.

The culture was analyzed by flow cytometry, and single cells were sorted into 96-well plates as described above. After overnight growth, 1 µl of cultures with visible growth were used as templates for rolling circle amplification (RCA) using TempliPhi DNA Sequencing Template Amplification Kits (Amersham Biosciences) with the recommended manufacturer protocol. RCA reactions proceeded in 10 µl volume at 30°C for 18 hours and then stopped by heating to 65°C for 10 minutes. 1U Shrimp Alkaline Phosphatase (SAP) was added to each reaction to dephosphorylate any remaining deoxynucleotides. Dephosphorylation proceeded at 37°C for 45 minutes followed by heat inactivation at 80°C for 10 minutes. 1 µl of the dephosphorylated RCA product was used as template for ½ volume Big Dye Terminator v3.0 sequence reactions with 4pmol of either forward or reverse M13 primers. Thermocycle conditions were 95°C for 2 min. and 25 cycles of 95°C 45 sec.,

50°C 30 sec.,60°C for 2 min.30sec. followed by 25 cycles of 95°C 45 sec., 50°C 30 sec., 72°C for 3 min. Sequence reactions were cleaned by isopropanol precipitation and run on ABI Prism 3700 DNA Analyzer (Perkin-Elmer). Sequences were analyzed and assembled using Phred/Phrap[16] and the assembly was viewed using Consed[17].

In the initial library, many inserts seemed to be derived from genomic *E. coli* DNA using the phenol/chlorophorm purification protocol. To overcome this problem, a second library was created using BAC DNA isolated with the Qiagen Large Construct Kit (Qiagen). The library was created in an identical manner with 2.5-5kb inserts. The same process was repeated for clone isolation by flow cytometry, culturing, amplification by RCA, and sequencing. The new protocol yielded clean DNA with minimal genomic contamination.

The 384 clones (768 sequencing reads) from the first contaminated library were individually screened using a Smith-Waterman comparison against the published *E. coli* genomic sequence. 309 sequence reads were found to have at least 100 bases of good quality sequence and no significant similarity to *E. coli* genomic sequence. These 309 sequences were combined with 416 sequence reads (208 clones) from the second library to perform the analysis.

In the end, the average length of Phred20 score sequence reads was approximately 350 bases per sequence. The final assembly (figure 21) consisted of 30 contigs, although many of these "contigs" contained only one or two clones and

---

[16] Ewing B., L. Hillier, M. Wendl, P. Green. 1998. "Basecalling of automated sequencer traces using phred." Genome Research **8**: 175-194.

[17] Gordon, D., C. Abajian, and P. Green. 1998. "Consed: A graphical tool for sequence finishing." Genome Research **8**:195-202.

probably represent small amounts of residual *E. coli* genomic DNA in the library. These orphan clones may also be the result of short sequence reads that could make sequence matching difficult. Of the 30 total contigs, 13 primary contigs were present that spanned 55kb of sequence out of a total BAC length of 59kb. This amount of successful assembly is expected given the level of coverage and the short average length of sequence reads. Sequence reads were evenly distributed throughout the assembly.

We analyzed the final assembly data to evaluate the insert lengths of our flow sorted clones. By looking at the locations of the forward and reverse reads within the final BAC assembly, we can estimate the distance between these reads. This distance is equivalent to the size of each insert. We analyzed 113 clones that had both forward and reverse reads residing on either of the two largest contigs. These contigs were 13.0kb and 7.3kb in size. The mean insert length was 2.95kb with a standard deviation of 858bp. This insert size mean is certainly within the 2.5-5kb range that was isolated during library construction. 6% of the inserts were smaller than 2kb in size. The final insert sizes may be slightly smaller than the starting library insert sizes. There may have been a measurable bias introduced toward smaller inserts during the bacterial culture phase prior to flow sorting. However, this effect does seem to be rather small.

## *E) Discussion*

The goal of this work was to demonstrate the isolation of individual bacteria containing cloned inserts by flow cytometry. We realized early in our experimentation that a positive fluorescence signal is necessary for flow cytometers to accurately sort only cells with inserts. Because *E. coli* cells vary widely in their cell-to-cell fluorescence levels, a sorted cell with low measured fluorescence levels could actually be a cell with a genotype that encodes for high levels of fluorescent protein expression. There are also a variety of weakly fluorescent background particles in culture media that must not be misidentified as cells to be sorted.

The initial pSE-GFP series vectors created a positive fluorescence signal when an insert was present. The cloning of an insert into the promoter region of the lac repressor gene allows for the transcription of GFP expressed from a separate lac promoter. Vectors without any inserts maintained low GFP levels. With inserts, cells exhibited a range of possible fluorescence values. A definitive boundary could never be established to separate the negative and positive populations. There was so much cell-to-cell variation in fluorescence levels of both populations that it became impossible to know for certain which cells contained an insert. This can be seen in figure 9.

To address the problem of cell-to-cell variability, we designed new vectors in the pBGFP series with two translationally fused fluorescent proteins. In the native vector, the fused BFP and GFP genes can be detected through their blue and green

fluorescences. When an insert is cloned in between the two genes, the production of the GFP half is prevented; this results in cells with only blue fluorescence. The two fluorescence profiles can be seen in the cells shown in figure 12. By using one fluorescent protein as a control for the other, the issue of fluorescence variability is mitigated. We can quantify the fluorescence ratio of BFP to GFP very accurately by flow cytometry, allowing us to differentiate the two distinct populations of cells. These two populations can be seen in figure 13. It should be noted that no reliance is made on expression levels, since we base sort decisions only on the ratio of the two fluorescences. As long as the expression levels are high enough to accurately quantitate the relative numbers of the two types of fluorophores, and accurate sort decision can be made.

The pBGFP series vectors allowed us to successfully separate insert containing and non-insert containing bacteria efficiently. However, in the end, a large number of sorted cells appeared to contain plasmids with rearrangements. Deletions and possible duplications were observed in up to 10% of clones grown in RecA minus host strains. These data will be presented in a later section entitled "RecA Independent Recombination".

To address the high frequency of recombination in pBGFP, we created the pGRFP series of vectors that contained a translational fusion of GFP and DsRed. The lack of sequence similarity between the two fluorescent proteins significantly decreased the frequency of recombination. Similar to pBGFP, the insert containing and non-insert containing bacteria in cultures could be easily distinguished under

fluorescence microscopy (see figure 16). The non-insert containing cells appear orange due to the combination of green and red fluorescences. The insert containing cells appear green due to the loss of DsRed translation. Flow cytometry was also performed on an Adenovirus-2 library cloned into pGRFP (see figure 17). The insert containing and non-insert containing populations could be easily separated according to the GFP:DsRed fluorescence ratios. Similar to pBGFP, very narrow distributions of fluorescence ratios were observed for each of the two populations.

The newly constructed pGRFP plasmid performed very well in clone selection experiments. By sorting bacteria in the GFP+/DsRed- population, 46 Adenovirus-2 DNA inserts were successfully amplified by PCR out of 48 clones selected for analysis (see figure 19). Only one clone potentially consisted of native pGRFP without an insert.

The use of pGRFP2 in a shotgun sequencing project clearly shows the usefulness of these vectors in real world applications. A partial assembly of the SU66E20 BAC was successfully completed. Some quality issues arose during the sequencing process. In particular, the final sequence analysis yielded short average sequence read lengths. We showed that many of these issues are related to the TempliPhi and sequencing protocols. Later experiments have shown that results can be significantly improved by taking steps to minimize evaporation during long incubations of the TempliPhi reaction. Nevertheless, a partial assembly of SU66E20 was created with 13 primary contigs that span 55kb of sequence from a 59kb BAC. This level of assembly is reasonable given the coverage as determined by the number

of clones sequenced and the average length of sequence reads. The use of cell sorting for clone selection did not generate an obvious bias toward a particular region of the BAC sequence. Furthermore, the size of cloned inserts was maintained roughly in the 2.5-5kb size range. There was only minimal bias toward smaller insert sizes that was introduced during the growth of transformed *E. coli* in culture prior to sorting.

We have showed the successful isolation of individual clones by single cell flow sorting. Isolated insert DNA has been shown to amplify well by PCR. Additionally, insert DNA has been used to partially sequence and assemble a 59kb BAC. These results show the robustness of the dual fluorescent protein method for isolating insert-bearing clones by flow cytometry.

With additional instrumentation, clone selection by sorting could be useful in high-throughput applications. In this work, all reactions were prepared by hand. Automated liquid handling could significantly improve the efficiency of the handling of clones after sorting. For instance, plates containing single sorted cells could be grown and then read in an instrument that could determine whether growth had occurred in each well. Such an instrument could work simply by measuring the amount of light absorption in each well from a passing laser beam. Sample handling robots could then harvest small amounts of these cultures for further amplification and processing. Use of TempliPhi rolling circle amplification reactions from sub-microliter volumes of culture could very quickly generate thousands of sequencing templates.

Further advances in single cell amplification technologies could make the described clone isolation strategy even more efficient. The ability to amplify inserts reliably from single sorted cells would eliminate the need for culturing and examining wells for growth. Despite the fact that our bacterial cells contain many hundreds of copies of our plasmids, single cell amplification of DNA can still be difficult. In our experience, PCR amplification is not robust enough to amplify inserts consistently from single cells. Some groups have had success in amplifying down to single molecules using rolling circle amplification techniques similar to the Templiphi kits used for this work.[18] In our own laboratory, we have shown some limited ability to use TempliPhi and GenomiPhi (Amersham Biosciences) RCA protocols for single cell amplification. There are also emerging new technologies that show promise for robust single cell amplification. One method of RCA developed by Tabor, et al.[19] is achieving sensitivities on the order of single cells.[20] Within several years, reliable DNA amplification from single cells will be available.

With the integration of cell sorting, DNA amplification technologies, and robotics, the described clone isolation strategy could truly become an efficient and practical method for isolating and processing large numbers of clones. Cell sorting has the potential to isolate individual clones at high speeds. New DNA amplification technologies might obviate the need for culturing clones at all, since plasmid DNA

---

[18] Lizardi, P.M., X. Huang, Z. Zhu et al. "Mutation detection and single-molecule counting using isothermal rolling-circle amplification." Nature Genetics **19**: 225-232.

[19] Kato, M., D.N. Frick, J. Lee, S. Tabor, C.C. Richardson, T. Ellenberger. 2001. "A complex of the bacteriophage T7 primase-helicase and DNA polymerase directs primer utilization." J. Biol. Chem. **276**:21809-20.

[20] Personal communication, Stan Tabor, Harvard Medical School, Boston, MA.

could be directly recovered from a single cell. Sequencing reactions could directly use amplified DNA, producing reaction products reactions that could be purified and analyzed with existing instruments.

The use of linear tape technologies as described in the "Introduction and General Background" chapter would allow us to sort in uninterrupted fashion into one long set of tubes. Given a conservative sort rate of 7000 clones per hour (approximately 2 per second), 168,000 clones could be isolated in 24 hours. This sort rate could easily double if we used two flow streams along with two separate carrier tapes. We could then isolate 336,000 clones per day using a single flow cytometer. Assuming the processing of forward and reverse sequence reads per clone and an average sequence read length of 500 bases, approximately 336 megabases of raw sequence data could be generated per day. At this rate, the human genome could be sequenced to 5X coverage in under 45 days. The methods described here could one day form the basis of efficient clone selection and sequencing capabilities that are not easily achieved today.

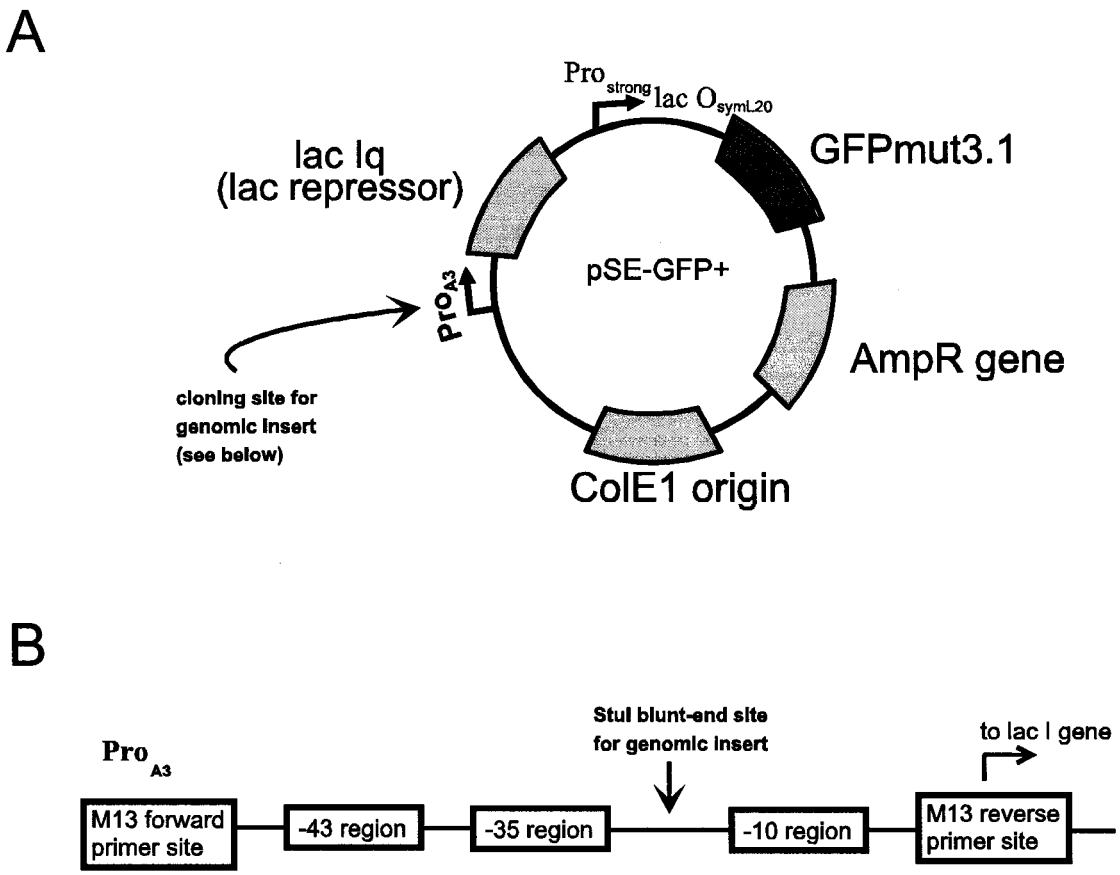
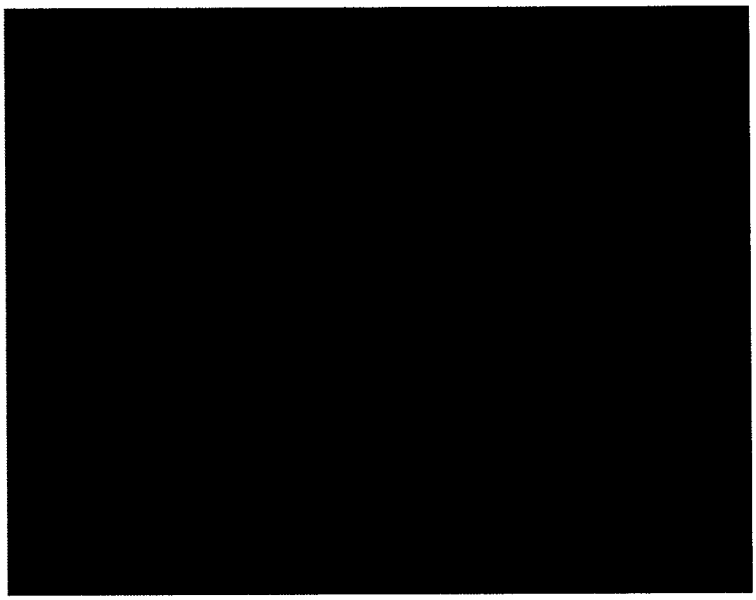



**Figure 7. Layout of pSE-GFP+ vector.** A) In the native pSE-GFP+ vector, the GFP gene is under the tight transcriptional regulation of a highly expressed lac repressor gene. This lac repressor is constitutively expressed from its own promoter on the plasmid. B) The promoter region of the lac repressor is shown in more detail. If an insert is cloned into a site between the -35 and –10 regions of the lac Iq promoter, expression of lac repressor will be shut down. This allows unfettered transcription of GFP. These bacterial clones can then be flow sorted based on their high level of GFP fluorescence. M13 forward and reverse primer sites are available flanking the insert site for PCR or sequencing.

A



B

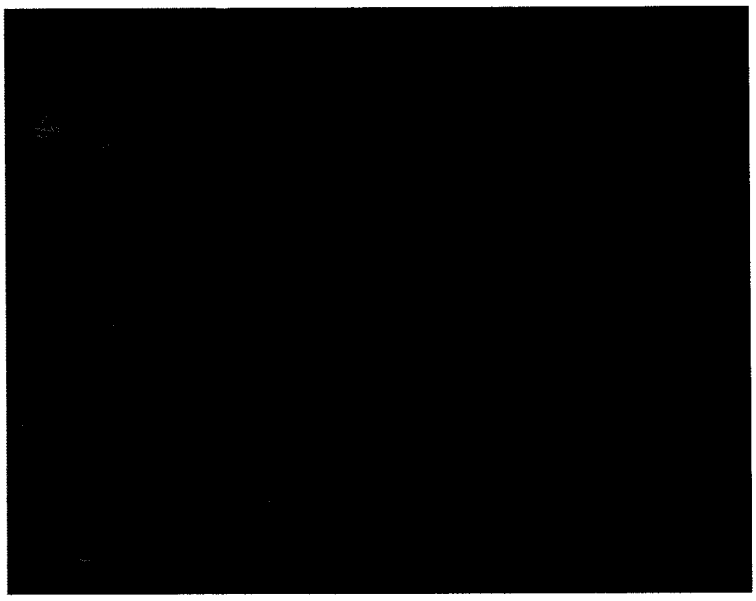


**Figure 8. Fluorescence microscopy images of E. coli, pSE-GFP+ series.** A) This is an image of a culture of *E. coli* containing native pSE-GFP+ vector. The expression of lac repressor prevents GFP production. B) This is an image of culture of *E. coli* grown under similar conditions but containing an insert cloned into the promoter of the lac repressor. Without lac repressor, GFP is produced unhindered.

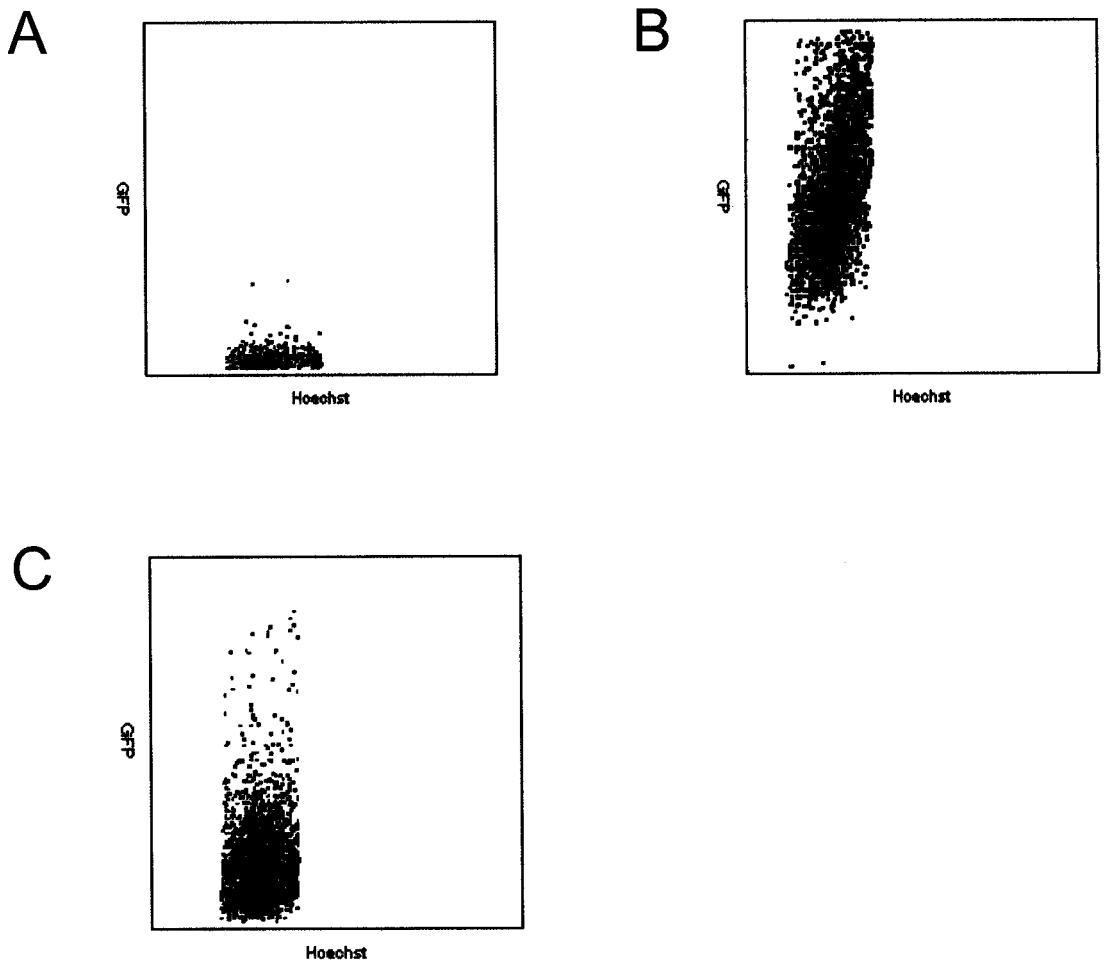



**Figure 9. Gated flow cytometry data from *E. coli* cultures containing pSE-GFP+.** Cells were stained with the viable DNA staining dye Hoechst 33342. Triggering on the flow cytometer was on this Hoechst parameter. The data were gated to encompass the major population of *E. coli* on the Hoechst axis. GFP fluorescence was plotted versus Hoechst fluorescence. A) *E. coli* cells containing native pSE-GFP+ have low levels of GFP fluorescence. B) *E. coli* cells grown from a single colony containing pSE-GFP+ plus a cloned insert. C) *E. coli* cells transformed with a constructed library of inserts in pSE-GFP+. There are both insert containing and non-insert containing bacteria in this sample. However, there is no clear boundary between the two populations.

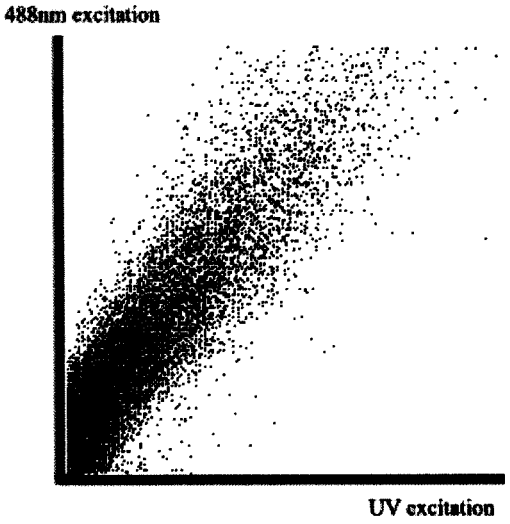
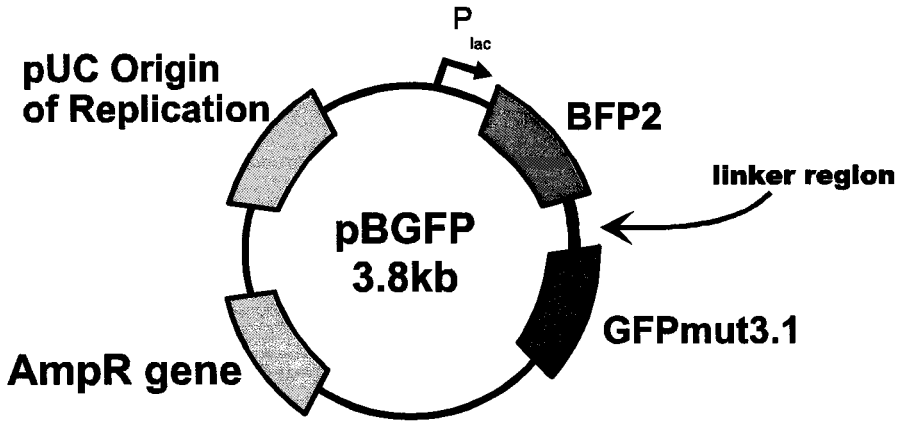



**Figure 10. Flow cytometry dot plot of stationary phase *E. coli* expressing GFP.** These cells were grown from a single colony and express GFP from a constitutive promoter. GFP fluorescence was measured as excited by two separate lasers: a 488nm beam and a multi-line UV beam. 488nm excited GFP fluorescence was plotted versus multi-line UV excited fluorescence. Note the large range of intensities of fluorescence present within the culture.

A




B

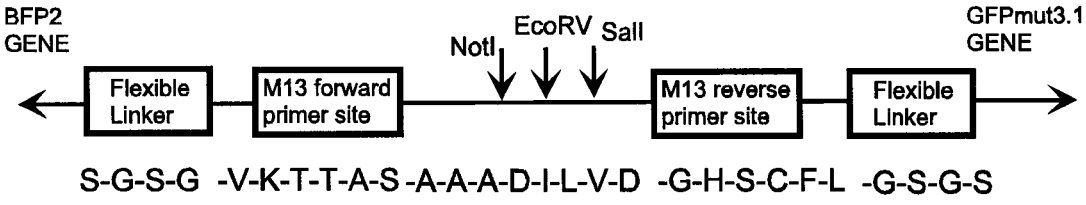



**Figure 11. Layout of pBGFP vector.** A) In the native vector, a fusion protein made up of BFP (blue mutant of GFP) and GFP is produced. If an insert is cloned into the vector between the two genes, stop codons within the insert will terminate translation of the GFP portion of the transcript. B) Layout of linker region between the BFP and GFP genes. A cloning site sits between universal M13 forward and reverse priming sites. A flexible linker is also included on each side to allow the BFP and GFP parts of the protein to assume orientations that are favorable for FRET to occur.

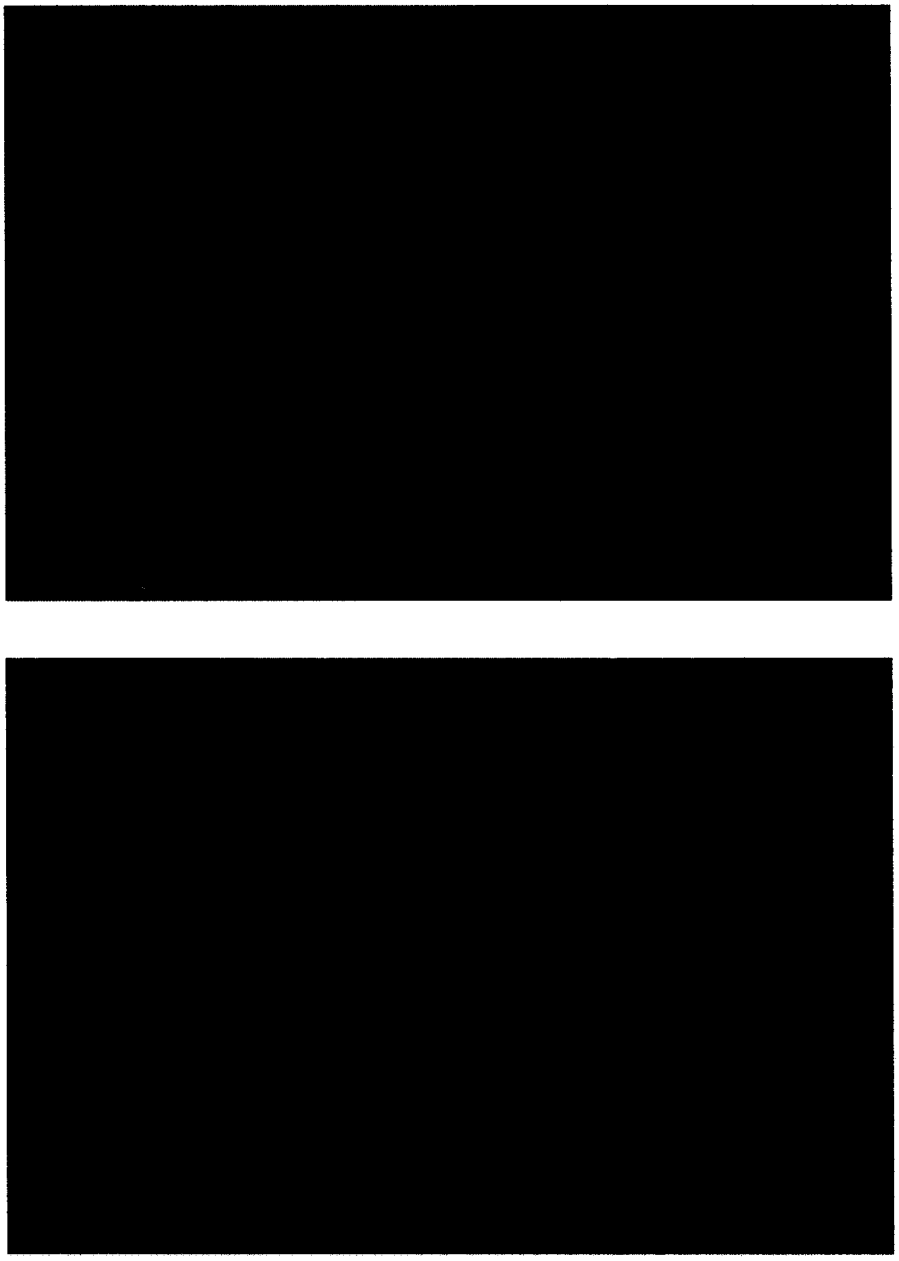


**Figure 12. Fluorescence microscopy images of E. coli, pBGFP series.** In the two images, cultures of *E. coli* containing pBGFP and a cloned Adenovirus-2 DNA library are shown. There are both insert containing and non-insert containing bacteria in this culture. The bacteria containing native pBGFP appear green due to the relative brightness of the GFP fluorophore and the partial FRET between the blue and green fluorophores. The insert containing bacteria appear blue due to the loss of translation of the GFP fluorophore. This color difference can clearly be seen in these images. Some of the bacteria in the lower image are out of focus.

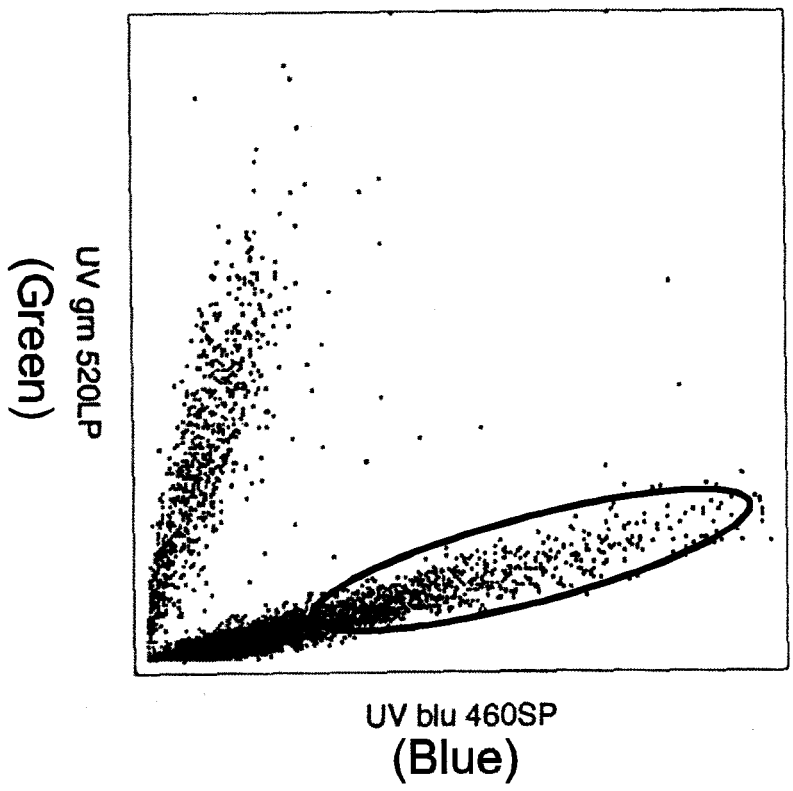



**Figure 13. Dot plot of E. coli culture containing pBGFP and cloned library.** An Adenovirus-2 library was cloned into pBGFP for this culture. Note the clear separation between the BFP/GFP containing cells versus the BFP only cells. Although there are large variations in expression levels, the ratios between the two fluorescences are well isolated to very narrow bands. Almost all of the cells lie on either of the two bands with very few cells in between. The blue circle indicates the BFP+/GFP- cells that presumably contain inserts. These cells can be sorted to isolate individual clones. The BFP+/GFP- cells do not lie perfectly on the x-axis due to some leakage of BFP fluorescence into the GFP detector.

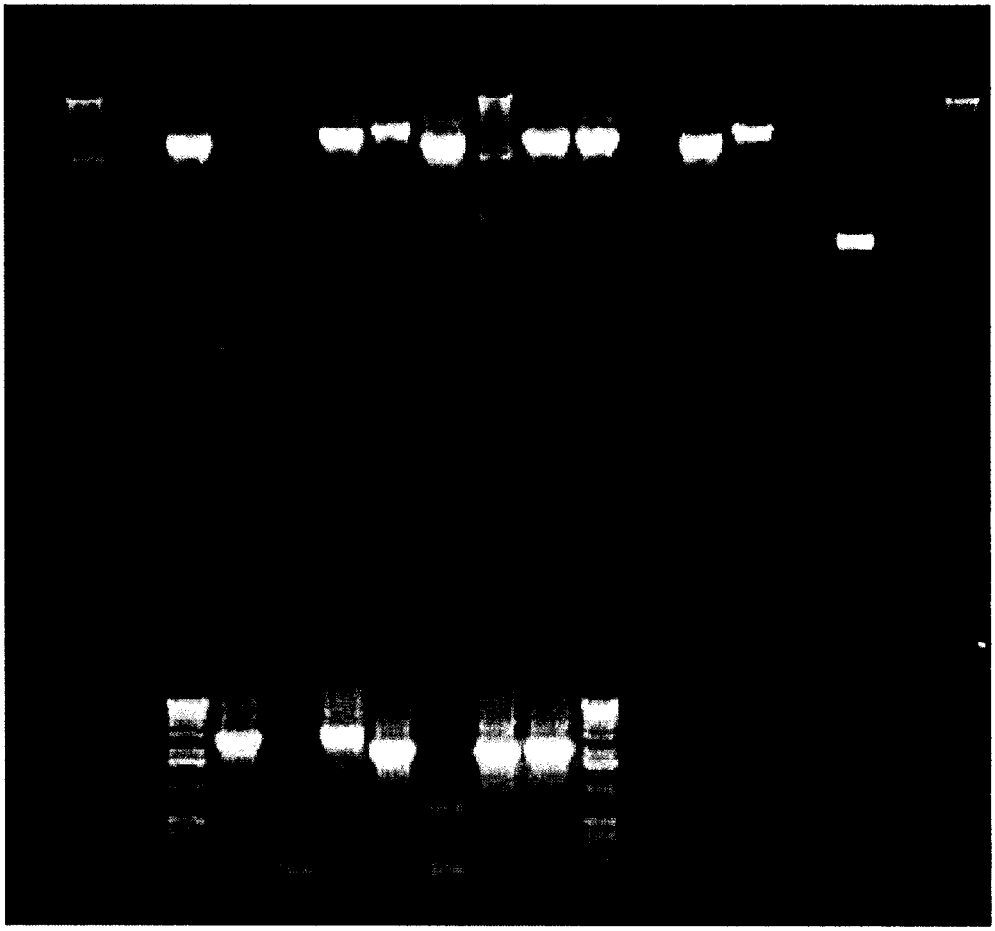


**Figure 14. Gel image of PCR amplified inserts after single cell sorting, pBGFP series.** Cultures were grown from single cells sorted from a culture containing an Adenovirus-2 library cloned into pBGFP. The cultures were used directly as templates for PCR amplification of cloned inserts. The last two PCR products (left of ladder) on the top portion of the gel are positive and negative controls from native pBGFP and water templates respectively. Of the 20 PCR reactions, 13 lanes show strong amplification of the cloned insert in the predicted 2-3 kb range. Seven PCR reactions resulted in weak or no PCR products.

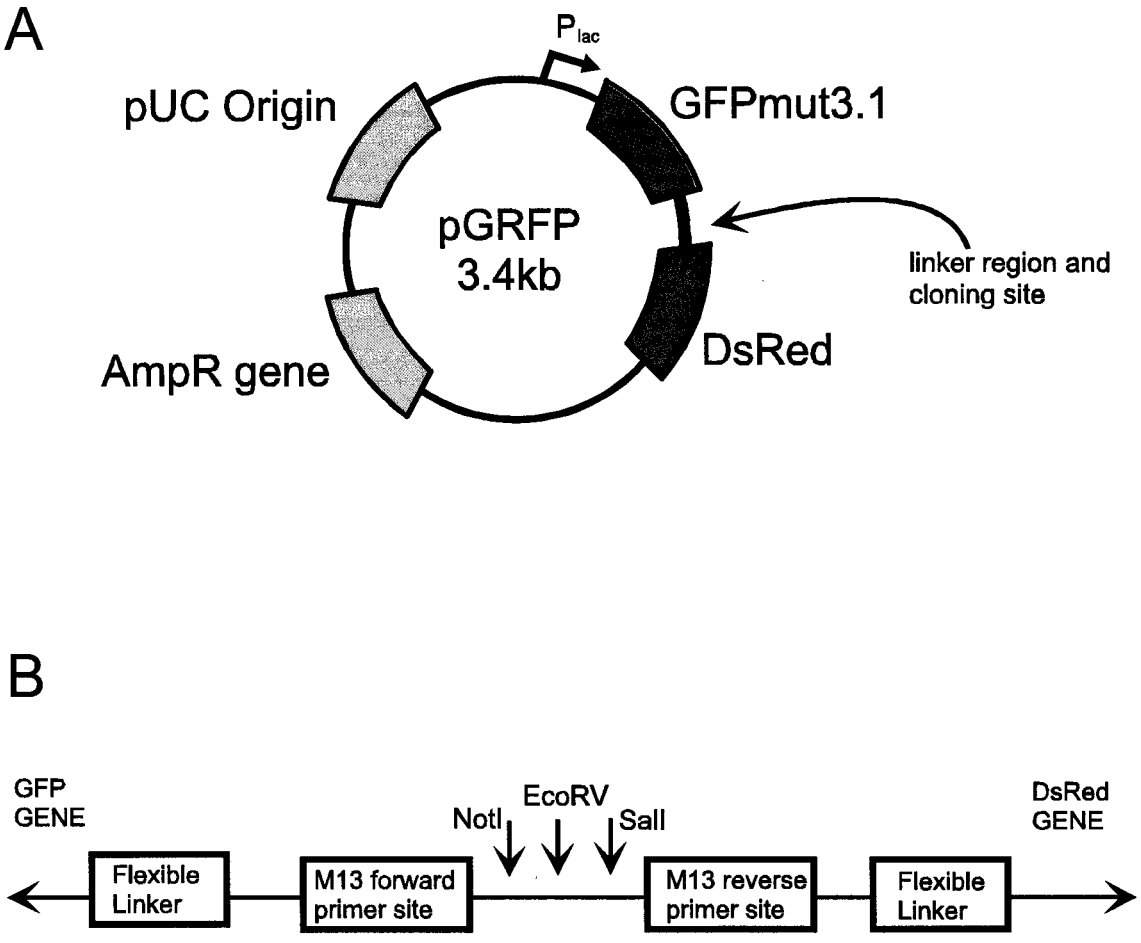



**Figure 15. Layout of pGRFP vector.** A) Similar to the pBGFP vector, a fusion protein consisting of two fluorescent proteins is produced by the native vector. Upon excitation by 488nm laser light, both GFP and DsRed fluorescence is elicited. If viewed under a microscope, bacteria have a yellow or orange hue. If an insert is cloned into the vector between the two fluorescent protein genes, stop codons within the insert will almost always prevent translation of the DsRed portion of the transcript. This results in exclusively green fluorescnce. B) Layout of linker region between the GFP and DsRed genes. A blunt end cloning site and two sticky end cloning sites sit between universal M13 forward and reverse priming sites. A flexible linker region is also included on each side to allow the GFP and DsRed proteins to assume orientations that are favorable for FRET to occur.

**Figure 16. Fluorescence microscopy image of E. coli, pGRFP series.** This image was produced of a culture of *E. coli* containing a library made from a mouse BAC subcloned into pGRFP. There are both insert containing and non-insert containing bacteria in this culture. The bacteria containing native pGRFP appear orange due to the excitation of both GFP and DsRed fluorophores. The insert containing bacteria appear green due to the loss of translation of the DsRed fluorophore. The color difference between these two populations is readily seen in this image.

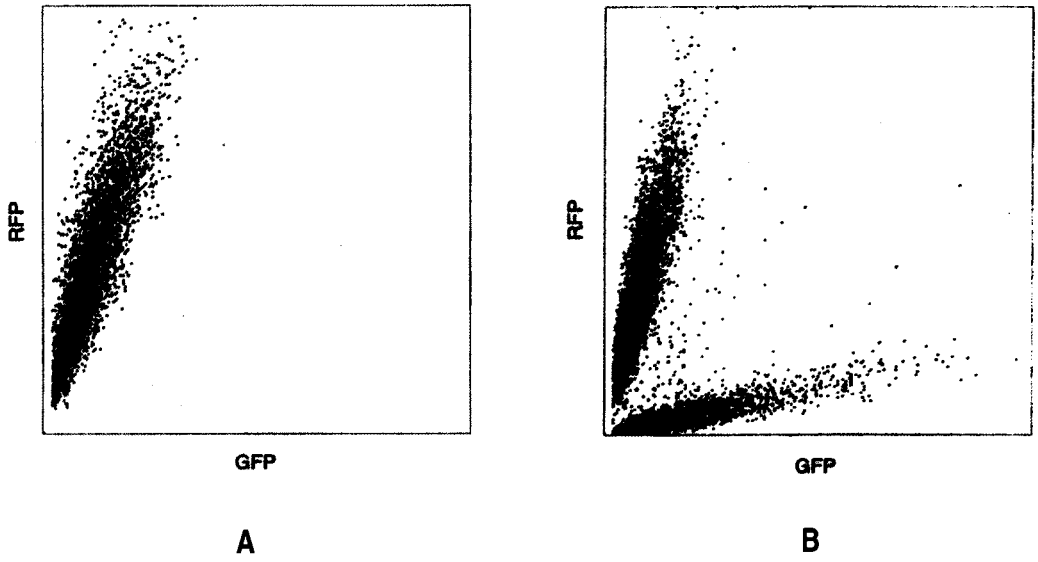



**Figure 17. Dot plot of E. coli culture containing pGRFP with and without inserts.** A) This dot plot shows a population of *E. coli* containing native pGRFP. All cells are seen expressing both GFP and DsRed. B) This dot plot shows an Adenovirus-2 library cloned into the pGRFP vector. A second population of cells that are GFP+/DsRed- appears due to loss of function of the DsRed half of the fusion protein. This assay distinguishes between insert containing and non-insert containing *E. coli*.

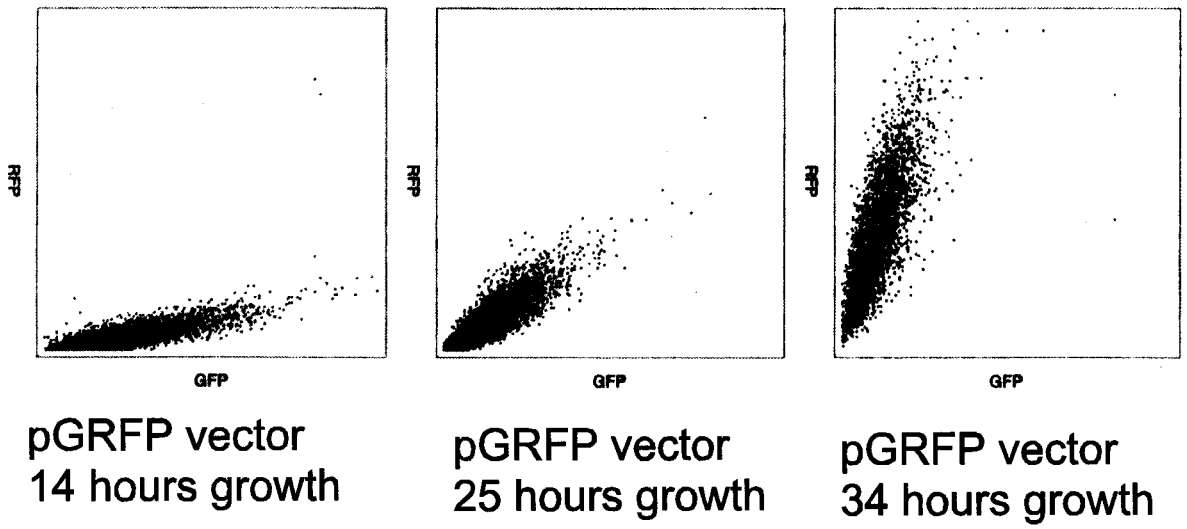



**Figure 18. Dot plots showing time course of fluorescence in pGRFP culture.** This culture was grown from *E. coli* transformed with native pGRFP plasmid. After14 hours growth, only green fluorescence is visible. The red fluorescence slowly develops over time until it reaches a level similar to what is seen in figure 13 for the GFP+/DsRed+ population. This highlights the greatly increased maturation time for DsRed versus GFP.

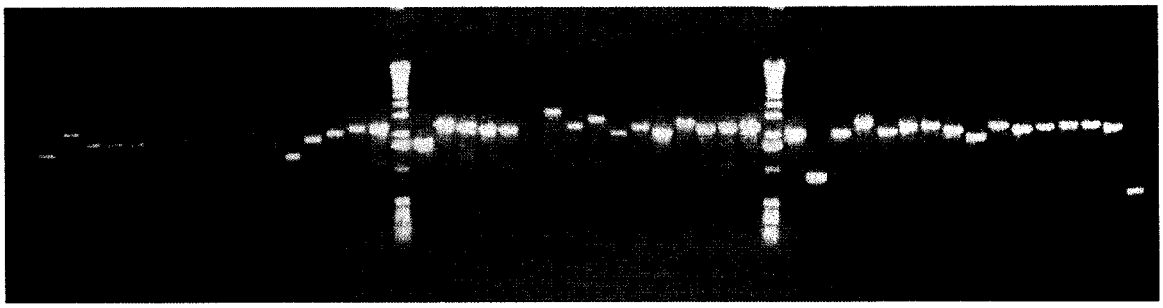


**Figure 19. Gel image of PCR amplified inserts after single cell sorting, pGRFP series.** Adenovirus-2 library inserts were cloned into pGRFP and transformed into DH10B *E. coli*. After 36 hours of growth, these cells were flow sorted into a 96 well culture plate, one cell per well. After growing overnight, confluent cultures were picked into PCR reactions optimized to amplify G-C rich regions. The resulting PCR products are shown above. Every well except one contained a single PCR product. The vast majority of these products were in the expected 1.5-3kb range. There was one 200bp product (in well 6) that appears to be from a small insert and not from amplification of the native pGRFP vector.

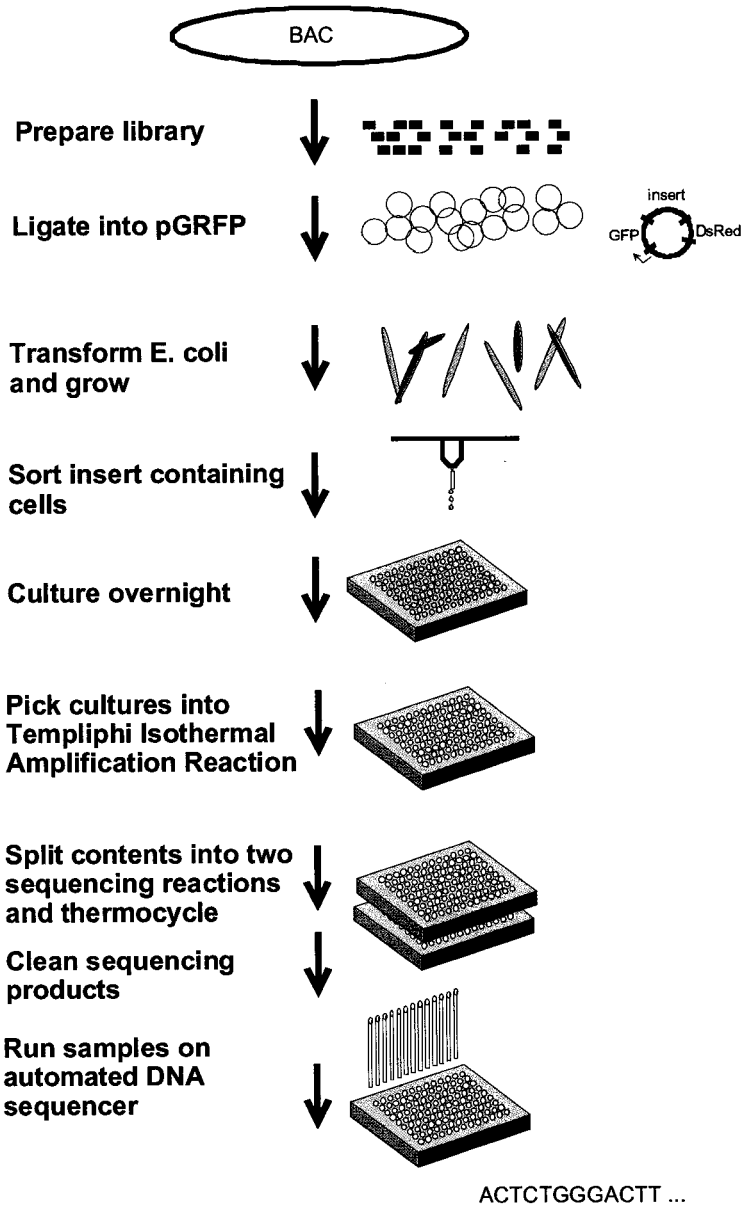



**Figure 20. Summary of BAC sequencing project with pGRFP2 vector.** The BAC was sonicated, end repaired, and blunt end cloned into pGRFP2. *E. coli* were transformed and grown. Single cells with the proper fluorescence profile were sorted and cultured. Plasmid amplification was performed with Templiphi RCA kit. These amplified products were the sequenced in the usual fashion. Sequence reaction products were cleaned and run on an ABI 3700 sequencing machine.

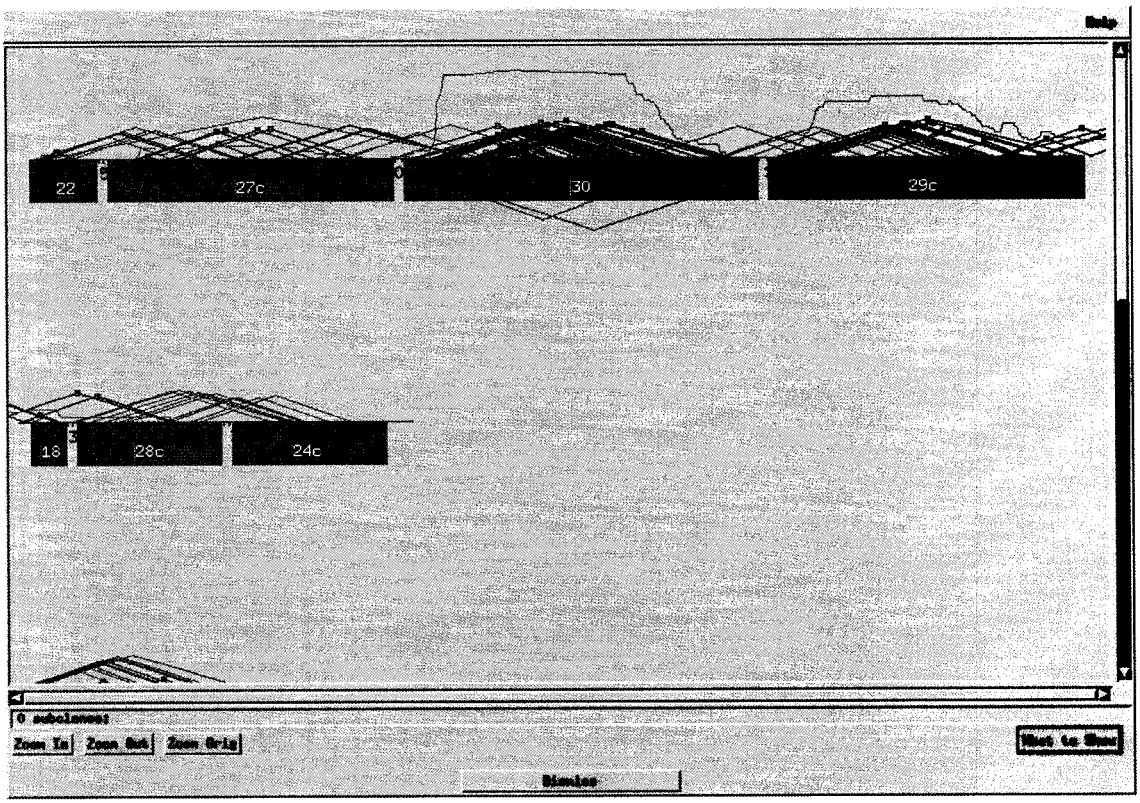



**Figure 21. Screen shot from Consed showing SU66E20 BAC assembly.** The largest contigs from the assembly are shown here. The total assembly yielded 13 primary contigs spanning 55kb out of a 59kb total BAC size. The many lines connect forward and reverse reads from the same clone.

# *Chapter 3. Protein Mutagenesis Project*

(In collaboration with Haiwei Guo and Larry Loeb, University of Washington, Seattle, WA)

## *A) Background – Mutation Tolerance*

Mutation of genetic material is a fundamental and necessary process of life. Mutations can be caused inadvertently from environmental or endogenous insults or through deliberate mechanisms. It is thought that the slow accumulation of mutations forms the basis of evolution. There is also strong evidence for the accumulation of mutations in cancerous cells.[1,2] Recently, it has become clear that one mode of innate cellular defense against retroviral infection is to hypermutate viral coding regions.[3,4]

Evolutionary studies and prior random mutagenesis experiments have led to the assertion that proteins are highly plastic against amino acid substitutions.[5,6] However, "highly plastic" connotes a subjective range of possibilities. Very few studies have investigated the sensitivity of proteins to mutation in a quantitative manner. Some

---

[1] Hanahan, D., and Weinberg, R.A. (2000). "The hallmarks of cancer". Cell **100**:57-70.
[2] Loeb, L.A., Loeb, K.R., and Anderson, J.P. (2003). "Multiple mutations and cancer." Proc Natl Acad Sci U S A **100**:776-781.
[3] Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S., and Malim, M.H. (2003). "DNA deamination mediates innate immunity to retroviral infection." Cell **113**:803-809.
[4] Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). "Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts". Nature **424**:99-103.
[5] Bashford, D., Chothia, C. & Lesk, A. M. (1987). "Determinants of a protein fold. Unique features of the globin amino acid sequences." J Mol Biol **196**: 199-216.
[6] Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). "Deciphering the message in protein sequences: tolerance to amino acid substitutions." Science **247**: 1306-10.

previous efforts that attempted to address this question were biased by local mutagenesis of small regions within proteins or by the correlation of relatively small numbers of mutations to protein function. Therefore, the tolerance of proteins to random mutations is still poorly understood.

Understanding protein robustness is critical for understanding processes such as evolution and carcinogenesis. More than thirty years ago, Maynard-Smith proposed that the set of functional mutant proteins that differ from the wild type by one residue must be very dense for evolution to be possible.[7] Measuring the robustness of various proteins could help confirm Maynard-Smith's hypothesis as well as increase our knowledge of how proteins are able to evolve over time.

In understanding carcinogenesis, the robustness of proteins is directly related to how prone organisms are to developing cancer. The loss of function of tumor suppressor genes is thought to be a critical step on the road of carcinogenesis. Knowing the level of robustness of cancer related genes could help correlate known rates of mutation accumulation to rates of carcinogenesis.

From a practical standpoint, knowledge of a protein's overall tolerance to mutation as well as the mutability of individual residues in the protein could provide valuable information for protein engineering. To create mutants of naturally occurring enzymes with altered functionality, many groups screen large mutant libraries looking for individual clones with the desired activity. Knowing a protein's overall mutation tolerance can help optimize the mutation frequencies of these mutant libraries. Higher

---

[7] Maynard-Smith, J. (1970). "Natural Selection and the Concept of a Protein Space." Nature **225**: 563-564.

mutation frequencies can potentially create many diverse clones that will approach a given target activity more quickly. However, higher mutation frequencies also lead to higher rates of protein inactivation. These two forces must be carefully balanced for directed evolution to succeed. In addition, knowledge of the tolerance of various amino acid positions to mutation can lead to effective libraries based on site-directed mutagenesis. For instance, most catalytic domains of an enzyme will be very sensitive to mutation. Selectively mutating these residues can lead to a faster convergence on the desired protein activity in some cases.[8]

In this work, we developed a method using the pGRFP series of vectors to rapidly quantify the robustness of various proteins to mutagenesis as well as measure the tolerance of every amino acid position to random mutagenesis. The robustness of proteins to mutations can be summed up as an "x-factor" that describes the probability that a random amino acid change will inactivate a protein. This "x-factor" is described in greater detail below. The mutation sensitivity of various amino acid sites indicates their importance to the protein's overall function. These amino acid sensitivities can be correlated with known structural information and evolutionary data about a particular protein. As a model human enzyme, we used the 33kD monomeric DNA repair enzyme 3-methyladenine DNA glycosylase (AAG). This enzyme was chosen, because it is relatively well studied and has an established selection system.

---

[8] Encell, L.P., D.M. Landis, and L.A. Loeb. (1999). "Improving enzymes for cancer gene therapy." Nat. Biotechnol. **17**:143-147.

### *B) Background - The 3-methyl Adenine DNA Glycosylase Enzyme*

The genomic DNA of all species continuously suffers damage from endogenous metabolites as well as exogenous agents. Left unchecked, these DNA lesions can cause deleterious mutations, aborted replication, or even cell death. To address the seemingly daunting task of finding and initiating repairs on a wide spectrum of possible DNA lesions, almost every organism has evolved mechanisms of base excision repair (BER). DNA glycosylases are specialized enzymes that initiate BER by recognizing and removing damaged bases. There are currently at least six known classes of DNA glycosylases for the repair of alkylated bases, deaminated, bases, oxidized bases, and mismatched bases.[9] The glycosylases cleave the N-glycosylic bond between the damaged base and the deoxyribose to leave an abasic site. The abasic site is then usually repaired by the action of several endonucleases, polymerases, and ligases that cut out and replace the base using the opposite strand as a template.

DNA glycosylases that recognize and initiate repairs on 3-methyladenine (3-MeA) have been found in a wide variety of organisms including bacteria, yeast, plants, mice, and humans.[10] The universal occurrence of this class of enzymes demonstrates its importance. Most of the known 3-MeA glycosylases also recognize and repair a variety of DNA lesions besides 3-MeA. For instance, the Tag enzyme in *E. coli* also repairs 3-methylguanine (3-MeG) in addition to 3-MeA. The AlkA

---

[9] Wyatt, M.D., J.M. Allan et al. (1999). "3-Methyladenine DNA Glycosylases: Structure, Function, and Biological Importance." Bioessays **21**:668-676.

[10] Memisoglu, A. and L. Samson. (1996). "DNA Repair Functions in Heterologous Cells." Crit. Rev. Biochem. Mol. Biol. **31**:405-447.

enzyme (also in *E. coli*) repairs a large number of potential lesions including, but not limited to, 3-MeA, 3-MeG, 7-methylguanine, $O^2$-methylthymine, $O^2$-methylcytosine, 7-chloroethylguanine, 7-hydroxyethylguanine, 3-chloroethylguanine and hypoxanthine (Hx). It can also remove normal bases at a low rate. The human version of this enzyme (human AAG) recognizes 3-MeA, 3-MeG, 7-MeG, 8-oxoguanine, and Hx. The *E. coli* AlkA gene and the human AAG gene do not share very much structural similarity. AlkA has two α-helical domains and one mixed αβ domain. It has an active site pocket lined with hydrophobic tryptophan and tyrosine residues that allow a diverse range of target substrates. The glycosylic bond seems to be cleaved by nucleophilic attack from a deprotonated water molecule.[11,12] The human AAG enzyme consists of a single mixed α/β domain containing a relatively flat DNA binding surface and a protruding hairpin loop that intercalates into the minor groove of the target DNA.[13] This intercalation of the hairpin structure displaces the target base and flips it into the active site of AAG. Similar to the *E. coli* AlkA enzyme, this active site is lined with aromatic amino acids. It has been demonstrated that the glycosylic bond is cleaved by an activated water molecule similar to the mechanism of AlkA.[14]

---

[11] Labahn, J., O.D. Scharer, et al. (1996). " Structural Basis for the Excision Repair of Alkylation-Damaged DNA." Cell **86**: 321-329.
[12] Yamagata, Y., M. Kato, et al. (1996). "Three-Dimensional Structure of a DNA Repair Enzyme, 3-Methyladenine DNA Glycosylase II, from *Escherichia coli*. Cell **86**:311-319.
[13] Lau, A., O. Scharer, et al. (1998). "Crystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanism for Nucleotide Flipping and Base Excision." Cell **95**:249-258.
[14] Lau, A., O. Scharer, et al. (1998). "Crystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanism for Nucleotide Flipping and Base Excision." Cell **95**:249-258.

In *E. coli*, Tag is a constitutively expressed 3-MeA glycosylase. AlkA, however, is induced upon the detection of DNA damage by alkylating agents through the Ada pathway. The Ada protein is a 39kDa that functions both as a transcriptional activator and a DNA-repair enzyme.[15] When it transfers a methyl group from an $O^6$-methylguanine or an $O^4$-methylthymine to one of its residues, the Ada protein becomes an efficient activator that upregulates the production of more Ada as well as the AlkA gene.[16,17]

Given these pathways, the *E. coli* strain MV1932 was created with mutations in both AlkA and Ada.[18] These cells are exquisitely sensitive to the actions of alkylating agents such as Methyl Methanesulfonate (MMS).[19] MMS causes a large number of lesions; 7-methylguanine is the most common lesion with 3-methyladenine occurring at 1/10 the frequency. MMS also rarely creates $O^6$-methylguanine, but these lesions are thought to be particularly lethal. Wildtype *E. coli* were relatively immune to the effects of .05% MMS, while less than .01% of MV1932 cells survived 60 minutes exposure to the same dose of MMS (see figure 22). Samson, et al. used this MV1932 strain to isolate the human AAG gene from a human cDNA library. The human cDNA library was expressed from bacterial promoters and transformed into MV1932.

---

[15] Shevell, D.E., B.M. Friedman, and G.C. Walker. (1990). "Resistance to Alkylation damage in *Escherichia coli*: Role of the Ada Protein in Induction of the Adaptive Response." Mutat. Res. **233**:52-72.
[16] Nakabeppu, Y. and M. Sekiguchi. (1986). "Regulatory Mechanisms for Induction of Synthesis of Repair Enzymes in Response to Alkylating Agents: Ada Protein acts as a Transcriptional Regulator." Proc. Natl. Acad. Sci. **83**: 6297-6301.
[17] Teo, I., B. Sedgwick, et al. (1986). "The Intracellular Signal for Induction of Resistance to Alkylating agents in E. coli." Cell **45**: 315-324.
[18] Volkert, M.R., D.C. Nguyen, and K.C. Beard. (1986). "*Escherichia coli* Gene Induction by Alkylation Treatment." Genetics **112**: 11-26.
[19] Samson, L. B. Derfler, et al. (1991). "Cloning and Characterization of a 3-Methyladenine DNA Glycosylase cDNA from Human Cells whose Gene Maps to Chromosome 16." Proc. Natl. Acad. Sci **88**: 9127-9131.

After several rounds of selection with MMS, a plasmid clone was isolated that conferred partial resistance to MMS (see figure 22).

Using this MV1932 strain, we performed early experiments to confirm that the human 3-Methyladenine DNA Glycosylase (AAG) is functional when expressed as a C-terminal fusion to GFP in pGRFP. Haiwei Guo (University of Washington, Seattle, WA) performed much of this early work. Cultures were grown from two clones containing AAG correctly oriented at the C-terminus of GFP (labeled AAG1 and AAG2). One control clone was also grown with AAG oriented in the backwards orientation (labeled GAA). After reaching late log phase, all three cultures were exposed to 0.2% MMS for 60 minutes. Cells were washed and plated cells onto carbenicillin plates at various dilutions. The relevant results after 24 hours growth are summarized below:

**Table 1: Protection of MV1932 cells with GFP-AAG fusion**

| Clone | Time in 0.2% MMS | Dilution | # Colonies | CFUs at 1:1 | Kill Ratio |
|---|---|---|---|---|---|
| AAG1 | 0 minutes | 1/100,000 | 175 | 17,500,000 | - |
| AAG2 | 0 minutes | 1/100,000 | 134 | 13,400,000 | - |
| GAA | 0 minutes | 1/100,000 | 135 | 13,500,000 | - |
| AAG1 | 60 minutes | 1/10,000 | 1197 | 11,970,000 | 1.46 |
| AAG2 | 60 minutes | 1/10,000 | 963 | 9,630,000 | 1.39 |
| GAA | 60 minutes | 1/100 | 119* | 11,900 | 1130 |

* Colonies were extremely small and slow growing, after an additional 34 hours, 268 colonies were counted

Comparing the kill ratios between the three clones, the properly oriented AAG gene fused to GFP complements the mutations in MV1932 well. The number of colony forming units (CFUs) decreased by only a factor of 1.39-1.46 after 60 minutes in 0.2% MMS. In contrast, the "GAA clone" had a 1130 fold decrease in CFUs. Therefore, proper AAG expression offers approximately 800-fold protection when cloned into the pGRFP vector.

Complementation of the MV1932 strain by functional AAG cloned into pGRFP forms the basis for our ability to catalog sets of mutations tolerated by AAG. By randomly mutagenizing AAG, selecting for functional mutants of AAG, and

sequencing them, we can quickly catalog the numbers, types, and locations of mutations that AAG can tolerate.

While we exploit MMS in this study for its cytotoxic effects, it is noted that MMS can also produce mutagenic lesions. Therefore, it is possible that some of the observed mutations could be caused through the process of selection. In order to address this issue, we sequenced the AAG coding region from sixteen clones containing wild-type AAG genes after cells underwent selection with 0.2% MMS. No mutations were observed from more than 10 kb of sequence. We are fairly confident from this data that the frequency of MMS mutation is very low in this system.

Because the crystal structure of human AAG has been solved, correlations between structure and function within AAG have been previously made.[20] In this work, we make our own correlations between mutation tolerant and mutation sensitive residues with what is known about AAG structure. We also correlate our mutation data with homology analyses performed by Lau, et al (see figure 23). In these analyses, human AAG was compared with homologues in mouse, rat, B. burgdorferi, B. subtilis, A. thaliana, and M. tuberculosis.[21] Some of the conserved bases correlate with important functional elements. For instance, Glu-125 that is thought to interact with water during the deprotonation step is absolutely conserved

[20] Lau, A., O. Scharer, et al. (1998). "Crystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanism for Nucleotide Flipping and Base Excision." Cell **95**:249-258.
[21] Ibid.

among the species analyzed. Interestingly, limited mutagenesis of human AAG at 17 residues has revealed that some substitution can occur even at sites that are highly conserved evolutionarily (see figure 24, Guo and Loeb, unpublished results). Random mutagenesis may have the potential to yield a greater diversity of amino acid substitutions than evolution has the ability to produce easily; conversely, one can say mutational mechanisms in evolution may be somewhat limited.

### *C) Mutagenesis, Selection and Sequencing*

We generated gene-wide mutations in the human AAG gene using two consecutive protocols for error-prone PCR. We used a combination of Mutazyme error-prone DNA polymerase[22] as well as Taq in the presence of $Mn^{+2}$ ions[23] to generate a wide spectrum of mutants of AAG. The published mutation spectrum of Mutazyme indicates a large bias toward mutating existing GC residues. 72.5% of all mutations generated are of GC residues versus 25.6% for AT residues.[24] A small number of mutations are insertions and deletions. The published mutation spectrum of Taq based mutagenesis with 640μM MnSO4 and 40μM dGFP was 17.8% of all mutations in GC residues versus 77.0% of mutations in AT residues. These figures assume that the template starts out balanced between A/T and G/C nucleotides. Looking at these mutation spectra, it appears that the two mutation protocols have

---

[22] Cline, J. and H. Hogrefe. (2000). "Randomize Gene Sequences with New PCR Mutagenesis Kit." Stratagies **13**:157-161.
[23] Clontech Laboratories. "Diversify PCR Random Mutagenesis Kit." Clontechniques **October 1999**: 14-15.
[24] Cline, J. and H. Hogrefe. (2000). "Randomize Gene Sequences with New PCR Mutagenesis Kit." Stratagies **13**:157-161.

complementary specificities for source bases of mutation. We therefore performed mutagenesis with both error-prone PCR protocols in series to create a more balanced spectrum of mutations in our libraries.

The theory in error-prone PCR is that the mutation frequency generated is proportional to the number of doublings in DNA quantity as given by the following formula.

$$Mutation\_frequency = (Polymerase\_error\_rate)x(Num\_doublings)$$

The number of doublings is given by the formula:

$$Num\_doublings = \log_2(amplification\_factor) = \log_2(\frac{PCR\_product\_quantity}{PCR\_template\_quantity})$$

Therefore, the amplification factor necessary to achieve a particular mutation frequency is given as:

$$amplification\_factor = 2^{\left(\frac{Mutation\_frequency}{Polymerase\_error\_rate}\right)}$$

The published polymerase error rate of Mutazyme is 0.310 mutations/kb·doublings.[25] In initial library construction attempts, we empirically

[25] Cline, J. and H. Hogrefe. (2000). "Randomize Gene Sequences with New PCR Mutagenesis Kit." Stratagies **13**:157-161.

measured the polymerase error rate of Taq under the specified conditions at 0.808 mutations/kb·doublings.

If we assume that mutation spectra are related to the nature of the polymerase and reaction conditions, then the mutation spectra should stay relatively constant for a given PCR protocol irregardless of the actual number of mutations generated. The ratio of A/T bases mutated to G/C bases mutated stays constant, and the actual value of the A/T and G/C mutation frequencies is directly proportional to the total mutation frequency. If we assume that the frequency of A/T base mutations and G/C base mutations is additive with two PCR reactions performed in series, we can estimate the final A/T and G/C base mutation frequencies for a balanced template given the total mutation frequencies in the Mutazyme and Taq reactions. These can be given with the matrices:

$$\begin{bmatrix} GC\_freq \\ AT\_freq \end{bmatrix} = \begin{bmatrix} 0.725 & 0.178 \\ 0.256 & 0.77 \end{bmatrix} \begin{bmatrix} Mutazyme\_freq \\ Taq\_freq \end{bmatrix}$$

Solving for Mutazyme_freq and Taq_freq, we get:

$$\begin{bmatrix} Mutazyme\_freq \\ Taq\_freq \end{bmatrix} = \begin{bmatrix} 0.725 & 0.178 \\ 0.256 & 0.77 \end{bmatrix}^{-1} \begin{bmatrix} GC\_freq \\ AT\_freq \end{bmatrix} \text{ OR}$$

$$\begin{bmatrix} Mutazyme\_freq \\ Taq\_freq \end{bmatrix} = \begin{bmatrix} 1.50 & -.347 \\ -.499 & 1.41 \end{bmatrix} \begin{bmatrix} GC\_freq \\ AT\_freq \end{bmatrix}$$

Therefore, given a particular desired A/T and G/C base mutation frequency, we can calculate the necessary number of amplifications for the Mutazyme and Taq reactions as follows:

$$\begin{bmatrix} Mutazyme\_doublings \\ Taq\_doublings \end{bmatrix} = \begin{bmatrix} 1/0.310 & 0 \\ 0 & 1/0.808 \end{bmatrix} \begin{bmatrix} 1.50 & -.347 \\ -.499 & 1.41 \end{bmatrix} \begin{bmatrix} GC\_freq \\ AT\_freq \end{bmatrix} \text{ OR}$$

$$\begin{bmatrix} Mutazyme\_doublings \\ Taq\_doublings \end{bmatrix} = \begin{bmatrix} 4.84 & -1.12 \\ -.618 & 1.75 \end{bmatrix} \begin{bmatrix} GC\_freq \\ AT\_freq \end{bmatrix}$$

We wanted to create libraries with low, medium, and high mutation frequencies with balanced mutations of G/C and A/T bases. We calculated the necessary number of doublings and amplification factors required for each protocol.

**Table 2: Calculated doublings and amplification factors for mutant libraries**

| library | Desired total mutation frequency | Desired A/T mutation frequency | Desired G/C mutation frequency | Desired Mutazyme doublings | Desired Taq doublings | Desired Mutazyme amp factor | Desired Taq amp factor |
|---|---|---|---|---|---|---|---|
| low | 4/kb | 2/kb | 2/kb | 7.44 | 2.26 | 174 | 4.79 |
| medium | 8/kb | 4/kb | 4/kb | 14.9 | 4.53 | 30,600 | 23.1 |
| high | 12/kb | 6/kb | 6/kb | 22.3 | 6.79 | 5,160,000 | 111 |

We performed the Mutazyme amplification first. We used the manufacturer recommended protocol adding Mutazyme buffer, 0.8mM dNTPs, 200 nM primers, and 2.5U Mutazyme polymerase in a 50 μl total reaction volume. We PCR amplified a ~1kb fragment containing the AAG gene from a 2.8kb plasmid vector. For the high library, we performed two Mutazyme PCR reactions in series to achieve the desired amplification factor. The starting and ending DNA quantities and amplification factors are summarized below.

**Table 3: Extent of amplification during Mutazyme PCR reaction**

| library | Starting template amount (2.8kb) | Final yield of 1.5kb gene fragment | Amplification factor | Doublings |
|---|---|---|---|---|
| low | 40ng | 4.99μg | 235X | 7.88 |
| medium | 232pg | 4.16μg | 33700X | 15.0 |
| High round 1 | 6.11ng | 4.05μg | 1250X | 10..3 |
| High round 2 | 876pg (1kb product) | 3.08 μg | 3519X (4,390,000X combined) | 11.8 (22.1 combined) |

We gel purified these PCR products and used these Mutazyme products as templates for the Taq reactions. The low library was PCR amplified with the high accuracy Pfu polymerase one time to generate enough product for the Taq reactions. The Taq reactions were performed according to the protocol for reaction condition 5 in the Diversify PCR Mutagenesis Kit manual from Clontech. Briefly, the reaction was performed with 1X Titanium Taq Buffer, 640 mM $MnSO_4$, 40 mM dGTP, 1X Diversify dNTP mix, 200nM primers, 1X Titanium Taq Polymerase in 50 μl total volume. The primers used were internal to the primers used in the Mutazyme reactions and yielded a 1.1kb product. The starting and ending DNA quantities and amplification factors are summarized below.

**Table 4: Extent of amplification during Taq PCR reaction**

| library | Starting template amount (1.5kb) | Final yield of 1.1kb gene fragment | Amplification factor | Doublings |
|---|---|---|---|---|
| low | 204ng | 895ng | 5.97X | 2.58 |
| medium | 62.2g | 1.50 μg | 32.9X | 5.04 |
| high | 13.6ng | 1.32 μg | 145X | 7.18 |

We PCR amplified these products with the highly accurate Pfu DNA polymerase using phosphorylated primers that amplified the AAG. The upstream primer 5'-GCCGCGGCCGCGAT-3' allowed the amplification of 4 codons at the N-terminus of the AAG gene. The downstream primer 5'-CCGCGGCGCGCTCGAGTC-3' adds 16 bases after the stop codon for AAG. The most 5' two bases were added so that if the PCR product was cloned into pGRFP2 in the backward direction, it would create a new BamHI site. This BamHI site can be used to reduce vectors with reversed inserts.

The cloned AAG gene was then propagated in *E. coli* strains. Haiwei Guo (University of Washington, Seattle, WA) performed the bacterial manipulations in this work. The Pfu amplified libraries were blunt-end cloned into the EcoRV site of pGRFP2 and transformed into DH10B cells and plated onto agar plates. After overnight growth, the plates were scraped and DNA harvested with a commercial maxi-prep kit. This DNA was cut with BamHI to eliminate pGRFP2 vectors with reversed AAG inserts. The remaining circular DNA was gel purified and re-transformed into MV1932 cells. Transformed cells were grown to confluence and

diluted 1:100 to mid-log phase. Cultures were treated with 0.2% MMS (Sigma) for one hour and drugs washed away. Treated and untreated cultures were serially diluted and plated on LB-carbenicillin in triplicate to calculate survival percentages.

We used these colonies grown on plates as a source for individual clones for further analyses. We initially attempted to isolate clones by single cell flow sorting similar to the methods employed in the sequencing project outlined in the "Sequencing Project" chapter. However, the MV1932 cells that were employed for these studies proved intolerant to flow sorting. Using the typical DH10B cells previously yielded growth in at least 50% of wells from single cells. The MV1932 cells typically yielded growth in <1% of wells when grown from single cells. Perhaps, the deficiencies in DNA repair mechanisms somehow prevent the recovery of these cells from the process of flow sorting. Cells are stressed by a variety of processes including injection into the flow stream at high pressures and brief illumination by an intense 488nm laser beam.

We therefore isolated clones from colonies on the agar plates for sequencing. All colony picking, DNA amplification, and sequencing steps were performed in cooperation with Haiwei Guo. From the plates originating from unselected cells, 24 colonies from the low library, 22 colonies from the medium library, and 47 colonies from the high library were picked into Templiphi reactions. DNA amplification with Templiphi proceeded similar to the procedures outlined in the "Sequencing Project" chapter. We sequenced with forward and reverse primers approximately 150bp away from the cloning site. Additionally, we used an internal primer that primes at base 108

of the AAG gene. The sequencing was performed with Perkin-Elmer Big Dye Terminator (BDT) 3.1 kits at 1/16X dilution in ½ volumes. Betaine was added at 1M concentration and 97 degree melting temperatures were used in cycling to overcome problems with GC-rich regions upstream of the start of the AAG gene.

The sequencing data for the unselected libraries were analyzed with Phred to perform the base calling. The Phred output files were submitted to Mutantman software (written by Juno Choe, currently unpublished) for high throughput comparison against the original AAG sequence. Mutantman is software that compares sequence and quality data from Phred to the starting consensus sequence. Phred quality data are used to determine whether a base sequence discrepancy is due to a real mutation generated during PCR or a product of a sequencing error. For our analyses, sequence data greater than Phred quality 20 were used. A modified Smith-Waterman alignment algorithm is used to align a set of sequence reads for a clone onto the consensus sequence. A wide variety of statistics can then be generated on the numbers, types, and locations of mutations on nucleotide and amino acid levels. Point mutations are detected as well as deletions, insertions, and early stop codons.

Using these methods, we looked at the mutation frequency and spectra of the three libraries as determined by 24 clones from the low library, 22 clones from the medium library, and 47 clones from the high library. The results are summarized below along with the desired mutation frequency and spectra that we set as our goal. We normalized mutation frequencies to account for the fact that AAG is GC-rich. All

mutation frequencies were normalized to predict what our mutagenesis protocol would have achieved with a gene with 25% A, 25% T, 25% G, 25% C bases.

**Table 5: Mutation spectrum of finished mutant libraries**

| library | Desired total mutation frequency | Desired AT freq / GC freq. | Total mutation frequency (normalized) | AT freq / GC freq. (normalized) |
|---|---|---|---|---|
| low | 4/kb | 1 | 3.41/kb | 1.29 |
| medium | 8/kb | 1 | 9.59/kb | 1.37 |
| high | 12/kb | 1 | 10.63/kb | 0.836 |

We achieved normalized mutagenesis rates that approach our desired goals. The low library has a mutation frequency very similar to our desired goal of 4 per kb. The medium library is slightly more than our desired goal, and the high library has slightly less than our desired goal. More importantly, all three libraries appear much more balanced than could have been achieved with Mutazyme or Taq protocols alone. The (AT freq)/(GC freq) ratio is fairly close to the desired value of 1 for all three libraries. In comparison, the Mutazyme protocol typically yields an (AT freq)/(GC freq) ratio of 0.353, and Taq yields a ratio of 4.33. This is good validation for our library construction protocols and calculations of the amplification factors required for the two complementary PCR mutagenesis reactions.

### *D) The x-factor*

### *D) The x-factor*

For the unselected libraries, Mutantman generated histograms for the clones with greater than 80% sequence coverage containing a given number of mutations. These histograms for all three libraries are shown in figure 25. A characteristic common to all mutagenic PCR libraries is that mutants show varying numbers of amino acid substitutions of mutations in an approximate gaussian distribution, which also broadens with increasing mutation frequency. An estimate of the percentage of clones in each library that have functional AAG was calculated by dividing the numbers of colonies originating from MMS-treated cultures by the numbers of colonies originating from unselected cells. The survival percentages are shown in figure 26. These values are important for quantifying the robustness of the AAG enzyme to mutation. In quantifying this robustness, we define the "x-factor" as the probability that a single random amino acid change will result in inactivation of the enzyme.

In its simplest form, the "x-factor" can be calculated from the fraction of functional clones in a library and the mean number of mutations per clone. However, this is a fairly crude method that does not account for the distribution of clones within the library. For instance, two libraries with the same mean number of mutations per clone can have very different survival fractions due to the distribution of mutations within each library. To better model the contribution to total survival fraction by various clones in given libraries, we have developed the following formula. The

assumption here is that individual mutations have a given probability (x-factor) of inactivating the enzyme that is independent of other mutations.

$$\sum_{n=0} f_n (1\text{-}x)^n = S$$

OR

$$f_0 (1\text{-}x)^0 + f_1 (1\text{-}x)^1 + f_2 (1\text{-}x)^2 + f_n (1\text{-}x)^n + \ldots = S$$

where:

$f_n$: fraction of unselected library with exactly n amino acid substitution(s)

n: number of amino acid substitutions

S: survival fraction of selected library compared to wildtype

x: probability of inactivating enzyme

The histograms in figure 25 constitute the $f_n$ values in the formula. The survival fractions given in figure 26 represent the S value for each library. Given this information, we calculated the x-factor to be 0.460, 0.395, and 0.383 from the low, medium, and high libraries respectively. We can further refine our x-factor by accounting for some mutations that are deletions or insertions instead of point mutations. It is assumed that these deletions and insertions result in frame shifted mutants that are always non-functional. We can adjust our x-factor as follows:

$$Xa = \frac{Xu - f}{1 - f}$$

where:

$x_a$: x-factor adjusted for insertion/deletion mutations

$x_u$: unadjusted x-factor calculated above

f: fraction mutations which are insertions or deletions

We can summarize the newly adjusted x-factors as:

**Table 6: Calculated x-factors for AAG from three mutant libraries**

| library | $x_u$: unadjusted x-factor | f: fraction ins/del | $X_a$: adjusted x-factor |
|---|---|---|---|
| low | 0.460 | 0.061 | 0.425 |
| medium | 0.395 | 0.100 | 0.328 |
| high | 0.383 | 0.054 | 0.347 |

Despite the widely varying survival fractions, we are able to calculate a relatively consistent x-factor for the AAG enzyme across our three libraries. This speaks to the validity of this method for estimating the mutation tolerance of a given protein. The actual x-factor for the AAG gene computed from a very large sample size would probably yield a value in the vicinity of 0.328-0.425.

### *E) Positional Tolerance to Mutation*

In order to analyze the nature of tolerated substitutions, we isolated and sequenced 274 active AAG mutants from the high library that survived treatment with MMS. Sequencing proceeded using similar methods described above for the clones that were untreated with MMS (see section C entitled "Mutagenesis, Selection and Sequencing"). All colony picking, DNA amplification, and sequencing steps were performed in cooperation with Haiwei Guo. The sequence data were analyzed by Mutantman and yielded a total of 920 amino acid substitutions. Figure 27 maps the mutations along the AAG primary sequence. The vertical colored bars above each residue indicate the frequency and types of tolerated amino acid substitutions at each position. Residues without bars indicate zero tolerated substitutions discovered. These sites are expected to include regions of important residues and motifs. Also, figure 27 shows the residues that are evolutionarily conserved among AAG homologues. Our mutation tolerance data correlate very well will conserved residues; many immutable residues in our data are evolutionarily conserved and perform essential roles in enzyme function. From the same data, we generated a 3D model to show the spatial orientation of residues that are both highly tolerant and highly sensitive to mutation. These models are shown in figure 28. As expected, these models show that regions of the AAG enzyme that reside in the core of the protein and are in the vicinity of the DNA substrate are highly immutable. The residues on

the outside surfaces of AAG show more flexibility in allowing mutations. Therefore, our data correlate well with evolutionary and crystal structure data.

More specifically, many correlations can be made between the mutability of individual amino acid positions and the crystal structure of AAG complexed to DNA.[26,27] Sidechains of Glu-125, Arg-182, and the Val-262 carbonyl oxygen directly contact an activated water molecule that is critical in the hydrolysis of the sugar-base glycosylic bond. All three positions are immutable in our study. Interestingly, Val-262 is not conserved evolutionarily despite its close contact with the water molecule. This suggests that our mutant selection scheme is even more effective at highlighting important residues than alignment with evolutionarily distant homologues.

Other amino acids important to AAG function include Tyr-162, Met-164, and Tyr-165. Tyr-162 projects from a surface β-hairpin and acts as a "nucleotide flipper" that moves the targeted nucleotide into the active site pocket. Met-164, and Tyr-165 assist in this base flipping mechanism by destabilizing the base pair adjacent to the flipped nucleotide. It has been shown that the Y162A mutant exhibits dramatically impaired glycosylase activity, while M164A and Y165A mutants display moderate impairment.[28] Correspondingly, in our study, the Tyr-162 position does not tolerate any changes, while positions Met-164 and Tyr-165 had only small numbers of

---

[26] Lau, A., O. Scharer, et al. (1998). "Crystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanism for Nucleotide Flipping and Base Excision." Cell **95**:249-258.

[27] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci **97**:13573-8.

[28] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci **97**: 13573-8.

substitutions, allowing two instances of M164I, one instance of M164R, and two instances of Y165F.

Within the substrate binding pocket, the flipped out base stacks between the aromatic sidechains of Tyr-127, His-136, and Tyr-159. Tyr-127 also stabilizes Glu-125 in the active site by donating a hydrogen bond. Using point mutants created at these three positions, Lau et al. demonstrated that Y127F has the least activity. The H136Q mutant had better activity, and Y159F had even better activity yet.[29] In our data set, Tyr-127 was correspondingly immutable. His-136 tolerated only one change to Tyr, and Tyr-159 was the most tolerant, allowing Phe and Asn changes. Our results correlate very well with the proposed action of these residues within the crystal structure of AAG and the limited point mutation data of Lau et al. The data further indicates that the relative frequency of mutation of particular residues correlates well with their relative importance to the proper functioning of AAG.

We also observed that many positions within AAG that were relatively intolerant to substitutions in our study were not evolutionarily conserved as we would normally expect. We explored the structural basis for their immutability and found three general themes: specific hydrogen bonding interactions, unique hydrophobic packing, and metal cofactor binding. With regards to hydrogen bonding interactions, Glu-116 interacts with Arg-118 and Glu-188 in a three-way hydrogen bonding interaction. Arg-261 provides a hydrogen pair partner to the evolutionarily conserved and immutable Asp-132.

---

[29] Ibid.

We propose that hydrophobic packing is another important reason why some residues of AAG are immutable. Gly-119 is at the core of a β fold less than 4.5A from Leu-184. No other residue besides glycine is able to fit in this core region. Similar packing constraints are observed with Leu184, which is less than 4.5A from Gly119 and less than 4.5A from the immutable Leu225. Cys167 is buried in a position less than 4.5A from Ile227 and very close to the Cα of Cys222 and Ala-183.

Metal cofactor binding can also be an important reason why we did not observe mutations in certain residues. Ser-171 was found to be immutable in our study, because it lies adjacent to a bound $Na^+$ ion that adds to the structural stability of the active site floor.[30] Interestingly, Ser-171 is the only residue that has a side-chain specific interaction with the $Na^+$ ion. $Na^+$ also contacts a water molecule and the main chain carbonyls of Met-149, Ser-172, Gly-174, and Ala-177.

There were also regions of AAG that were found to be highly mutable. Approximately 90 residues at the N-terminus of AAG had a high frequency of mutation. This result is not surprising, since it has been reported that the non-conserved 1-79 amino acids can be deleted without significantly affecting *in vitro* enzyme activity and DNA binding specificity.[31,32] The N-terminal region contains a protein-protein interaction domain with hRad23A and -B which can serve as

---

[30] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci **97**: 13573-8.

[31] O'Connor, T. R. (2000). "3-Methyladenine-DNA glycosylase (MPG protein) interacts with human RAD23 proteins." J Biol Chem **275**:28433-8.

[32] Roy, R., Biswas, T., Hazra, T. K., Roy, G., Grabowski, D. T., Izumi, T., Srinivasan, G. & Mitra, S. (1998). "Specific interaction of wild-type and truncated mouse N-methylpurine-DNA glycosylase with ethenoadenine-containing DNA." Biochemistry **37**:580-9.

accessory proteins for DNA damage recognition.[33] It is interesting to note that residues Gln23, Pro45, Ala54, Glu59, and Gly63 only had one instance of mutation each, even within a highly substitutable region. This possibly reflects requirements to preserve overall proper folding of AAG. Hence, the N-terminus acts as an internal positive control for mutability. In the crystal structure, residues 80-81, 200-207, 249-254, and 296-298 are inferred to be disordered segments due to their lack of electron density.[34] Accordingly, the frequency of mutation at these sites was found to be very high.

We also observed regions with a striking pattern of alternating immutable and highly mutable residues. This is seen in the β4 (164-172) strand. We hypothesize that this is due to the fact that alternating residues face in opposite directions. The residues facing toward the active site, Cys-167, Asn-169, and Ser-171, are involved in crucial interactions described earlier. These residues are therefore immutable. Residues with side chains pointing into the hydrophobic core tend to tolerate substitutions well, but only if the residues are substituted with another hydrophobic amino acid.

---

[33] Miao, F., Bouziane, M., Dammann, R., Masutani, C., Hanaoka, F., Pfeifer, G. & O'Connor, T. R. (2000). "3-Methyladenine-DNA glycosylase (MPG protein) interacts with human RAD23 proteins." J Biol Chem **275**:28433-8.

[34] Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci **97**: 13573-8.

## *F) Discussion*

It has been hypothesized in the past that proteins tolerate a wide range of single amino acid substitutions.[35,36] We wished to test this hypothesis and to establish a quantitative measure of the tolerance of enzymes to random substitutions. We define the concept of an "x-factor" that defines the probability of inactivation of any protein due to any given random amino acid change along the primary sequence. In this study, we estimated the x-factor of AAG at 32.8-42.5%. This is the first calculation of this type that we are aware of applied to a protein.

Our results beg the question of whether our estimated x-factor can be applied more generally to other proteins. At one extreme it can be argued that proteins perform vastly different functions, and the varying requirements would impose vastly different mutation tolerances on each individual protein. For instance, we can expect histones to be relatively sensitive to mutations, because they have a wide variety of DNA and protein binding constraints. Other genes such as antibody Fv regions have evolved for a high degree of tolerance to mutation. Arguing for relatively homogeneous x-factors, proteins face many of the same requirements such as having to properly fold into globular structures that are soluble in the surrounding medium. If a relative narrow range of x-factors does exist for many proteins, we can eventually

---

[35] Bashford, D., Chothia, C. & Lesk, A. M. (1987). "Determinants of a protein fold. Unique features of the globin amino acid sequences." J Mol Biol **196**: 199-216.

[36] Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). "Deciphering the message in protein sequences: tolerance to amino acid substitutions." Science **247**: 1306-10.

estimate this x-factor range by studying multiple proteins. In the meantime, we can attempt to corroborate our x-factor estimate for AAG with a variety of previous studies.

It has been suggested that the primary constraints on protein function include the ability to properly fold into soluble, globular structures and to maintain secondary structure.[37] In remaining soluble, globular proteins typically bury hydrophobic side chains within the protein while exposing hydrophilic side chains to the solvent side. For secondary structure such as α-helices and β-strands, the pattern of hydrophilic and hydrophobic residues has also been shown to be critical.[38,39] Kamtekar et al. has suggested a "binary code" model in which new proteins could be designed by maintaining the pattern of hydrophobic and hydrophilic residues without constraining the exact types of hydrophobic and hydrophilic residues used.[40] They tested this model with a four α-helix bundle; the majority of binary code homologues of the original sequence maintained their secondary structure and solubility. These results are collaborated by our own results in figure 27 that show that most tolerated mutations are conservative in nature.

Assuming that one deviation from Kamtekar et al.'s binary code model will result in loss of solubility or secondary structure, we can roughly estimate the

---

[37] Beasley, J.R. and M.H. Hecht. (1997). "Protein Design: The Choice of *de Novo* Sequences." J. Biol. Chemistry **272**:2031-2034.

[38] Xiong, H., B.L. Buckwalter, H.-M. Shieh et al. (1995). "Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides." Proc. Natl. Acad. Sci. **92**:6349-6353.

[39] West, M.W. and M.H. Hecht. (1995). "Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins." Protein Sci. **4**:2032-2039.

[40] Kamtekar, S., J.M. Schiffer, J.M. Xiong et al. (1993). "Protein design by binary patterning of polar and nonpolar amino acids." Science **262**:1680-1685.

probability that a single amino acid substitution caused by a point mutation will inactivate a protein. We can divide amino acids into hydrophobic residues, hydrophilic residues, prolines, acidic residues, and basic residues. A substitution occurring at a hydrophobic residue as caused by a point mutation has a 47.2% probability of becoming a hydrophilic residue, proline, acidic residue, or basic residue. Similarly, hydrophilic residues have a 68.7% probability of becoming one of the other types. In vertebrates, hydrophobic residues are observed in 40.1% of all amino acids, and hydrophilic residues occur at 29.0% of positions. A weighted average of these probabilities results in an estimated x-value of 0.563. This is somewhat larger than our estimated x-factor for AAG of .328-.425. However, the 0.563 x-factor value is likely an overestimate due to our tight constraint that all residues must absolutely remain within their amino acid class for a protein to maintain function. In other words, proteins are probably more flexible in accommodating mutations than the "binary code" model may suggest.

We also extend our analyses to the small number of published large-scale protein mutagenesis and functional studies. Markiewicz et al. examined amino acid substitutions at every residue across 90% of the *E. coli* lac repressor gene.[41] We reanalyzed their data and calculated the percentage of intolerant amino acid changes that resulted in totally inactive or temperature-sensitive phenotypes out of ~4049 total

[41] Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence." J Mol Biol **240**:421-33.

changes made. We obtained an x-factor for the lac repressor gene of 34%. This correlates remarkably well with our own results for the human AAG.

We also reanalyzed the results from Kawate et al. in which a single mutagenic PCR library of human thymidylate synthase was selected for activity and resistance to 5-fluorodeoxyridine.[42] This revealed a calculated x-factor of 22%. This represents a value slightly lower than the value we calculated for AAG. This may indicate that thymidylate synthase is a more robust enzyme. Alternatively, some sample bias could be inherent in this x-factor determination due to the small numbers of actual mutant clones sequenced in the Kawate et al. study.

The possibility exists that our calculations for the x-factor of AAG were lowered somewhat artificially by the presence of the most N-terminal 79 amino acids. This region has been shown to be unnecessary for enzymatic activity. These functionally unimportant regions do not occur in the majority of proteins that have been studied. In our study, 36% of all tolerated substitutions were found in these first 79 amino acids as compared to the 26% percent expected by chance. It is important, however, to note that this region still needs to abide by principles of proper protein folding as the rest of the enzyme. It is possible for mutations to occur in the first 79 amino acids that would lead to global protein misfolding. More experiments need to be performed for a wide variety of proteins from various organisms to determine whether our calculated x-factor for AAG is truly applicable to proteins in general.

---

[42] Kawate, H., D.M. Landis, and L.A. Loeb. 2002. "Distribution of mutations in human thymidylate synthase yielding resistance to 5-fluorodeoxyuridine". J. Biol. Chem. **277**: 36304-36311.

Organisms likely have evolved robustness against perturbations, from the organismal level to the molecular level.[43] In terms of amino acid choice, this principle is most apparent in the third position degeneracy of codons. The redundancy principle is also evident in transitions between amino acids as accessible by one nucleotide change. On average 5.7 types of amino acid substitutions are possible from any give amino acid residue. Single base changes often yield conservative amino acid substitutions.[44] In this study, the mutagenesis technique of nucleotide changes may be a better simulation of natural processes compared with other published methods such as alanine scanning that replace entire codons. In fact, out of the ten nonsynonymous AAG single nucleotide polymorphisms (SNPs) found so far in non-diseased human populations[45], three (Pro64Leu, Thr199Ala, Ala258Val) have been independently discovered in our tolerated substitutions database. The other seven SNPs have most likely been missed due to chance, and may be detected if more surviving mutants are sequenced. None of the ten naturally occurring SNPs are found at residues found to be immutable in our study. Therefore, SNP evidence correlates strongly with our map of mutability throughout the AAG gene.

One interesting question in evolution is the rate of introduction of deleterious alleles into genomes. For humans, it is estimated that the mutation rate per effective coding genome is 1.6 changes every generation. This may include base pair

---

[43] Edelman, G. M. & Gally, J. A. (2001). "Degeneracy and complexity in biological systems." Proc Natl Acad Sci **98**: 13763-8.
[44] Miyazaki, K. and F.H. Arnold. (1999). "Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function." J. Mol. Evol. **49**:716-20.
[45] http://www.genome.utah.edu/genesnps/cgi-bin/frame.cgi?gene_id=274

substitutions, insertions, deletions, and larger changes.[46] Using this figure and estimating the average x-factor for proteins at 35%, approximately 0.5 genes will be adversely affected every generation. However, this is likely an underestimate, since frame shift mutations and larger rearrangements will more readily inactivate genes. In the human species and other organisms, we do not directly observe a high frequency of gene inactivation. This is likely due to the large amount of redundancy built into biological systems. The diploid nature of our chromosomes compensates for most genetic defects. Even when both copies of a particular gene are inactivated, an observable phenotypic change may not result due to complex compensatory mechanisms provided by other gene products. Even if a phenotypic change results, it will not be observed generally if it results in death during development.

Organisms facing high mutation rates, such as HIV, may have proteins that have evolved some level of tolerance for amino acid substitutions. In this case, we would observe an x-factor significantly lower than for proteins found in other organisms and maybe even other viruses that do not experience the same frequency of mutation. It would be of interest to examine x-factors for proteins from various organisms to see if certain species have selected for a higher tolerance to mutation.

In this study, we have produced an analytical approach that provides a relatively expedient method for obtaining quality data about the mutability of amino acid sites throughout a protein. As a general approach, it lends itself to high throughput

[46] Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). "Rates of spontaneous mutation." Genetics **148**:1667-86.

automation using high speed sequencing and streamlined computational tools such as the Mutantman program used in this study. We were unable to utilize single cell flow sorting in this application due to the sensitivity of MV1932 cells. However, one can imagine that there are thousands of other proteins that could be analyzed using a flow cytometry approach. After mutation of a particular protein, the library can be assessed according to a selection or screening method. Using a selection approach such as in this study, surviving cells could be cultured and quickly isolated into single cells using a flow sorter. With a screening method, a fluorescent tag could be utilized to indicate whether the mutated protein is functional or not. These fluorescent tags could be read quickly on a flow sorter and used to separate single cells. In the future, it may be possible using the described methods to assess tolerance of mutations from a variety of proteins in a high throughput manner.

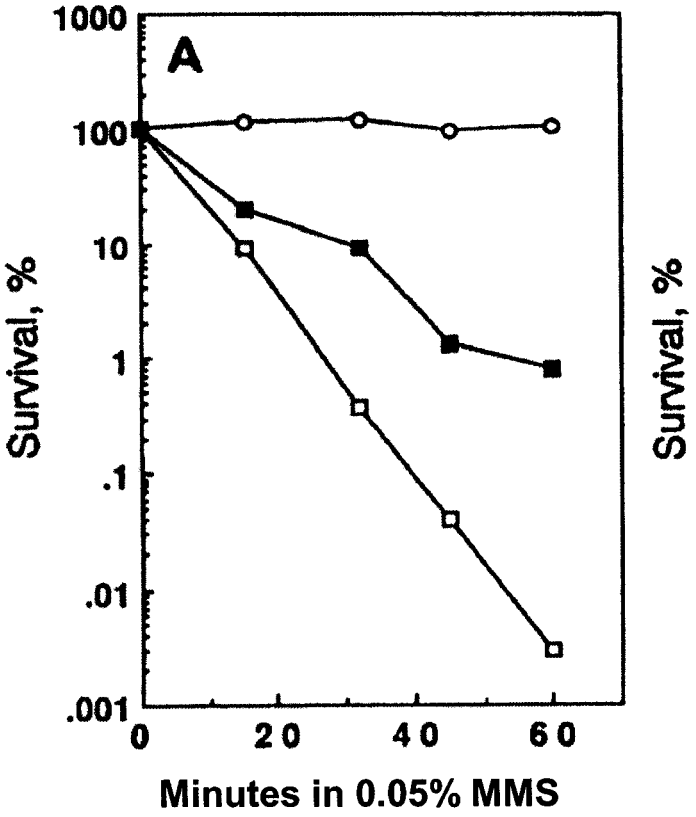



**Figure 22. Kill curve for the MV1932 strain exposed to 0.05% MMS over time.** The open circles represent wild type (AB1157) *E coli*. The open squares represent the MV1932 (alkA tag mutants). The closed squares represent the partially rescued MV1932 with a plasmid expressing the human AAG cDNA.

Figure reproduced from Samson et al. 1999.

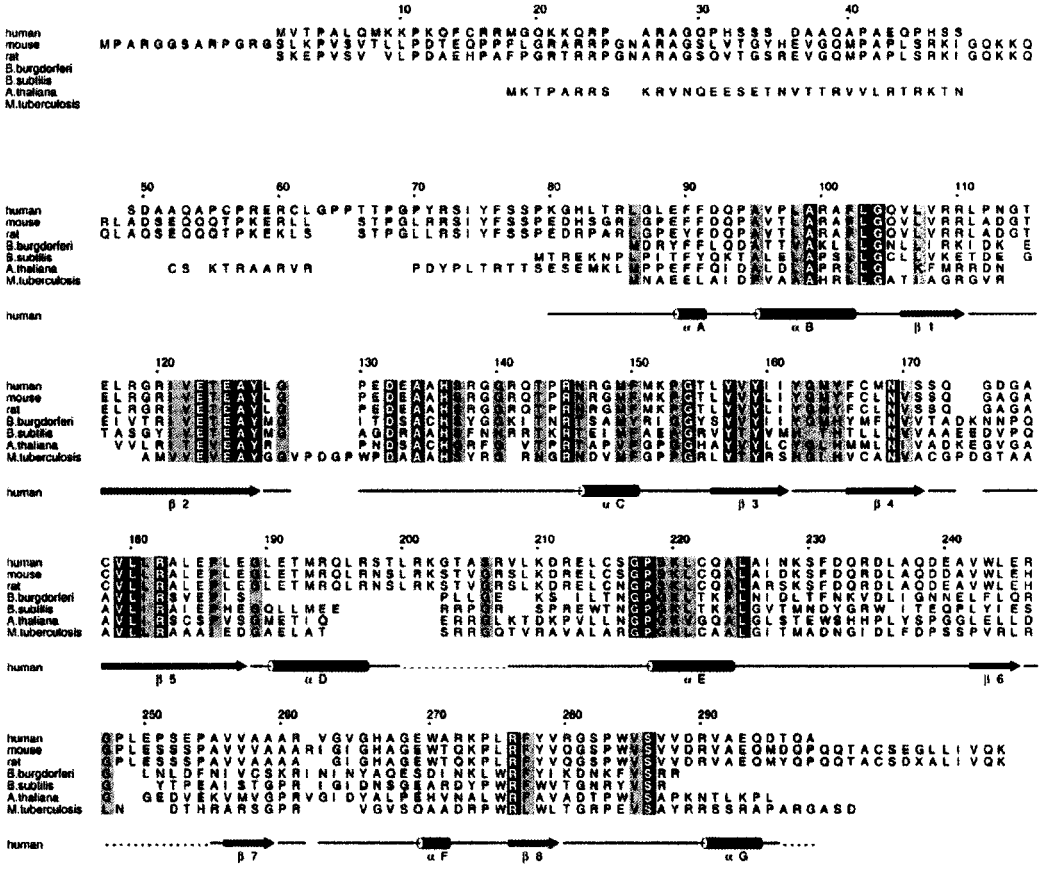



**Figure 23. Aligned sequence of human AAG and related DNA glycosylases.** Structural elements including α helices and β sheets are marked below the alignment. Absolutely conserved residues are marked in purple. Partially conserved residues are marked in orange.

Figure reproduced from Lau et al. 1996.

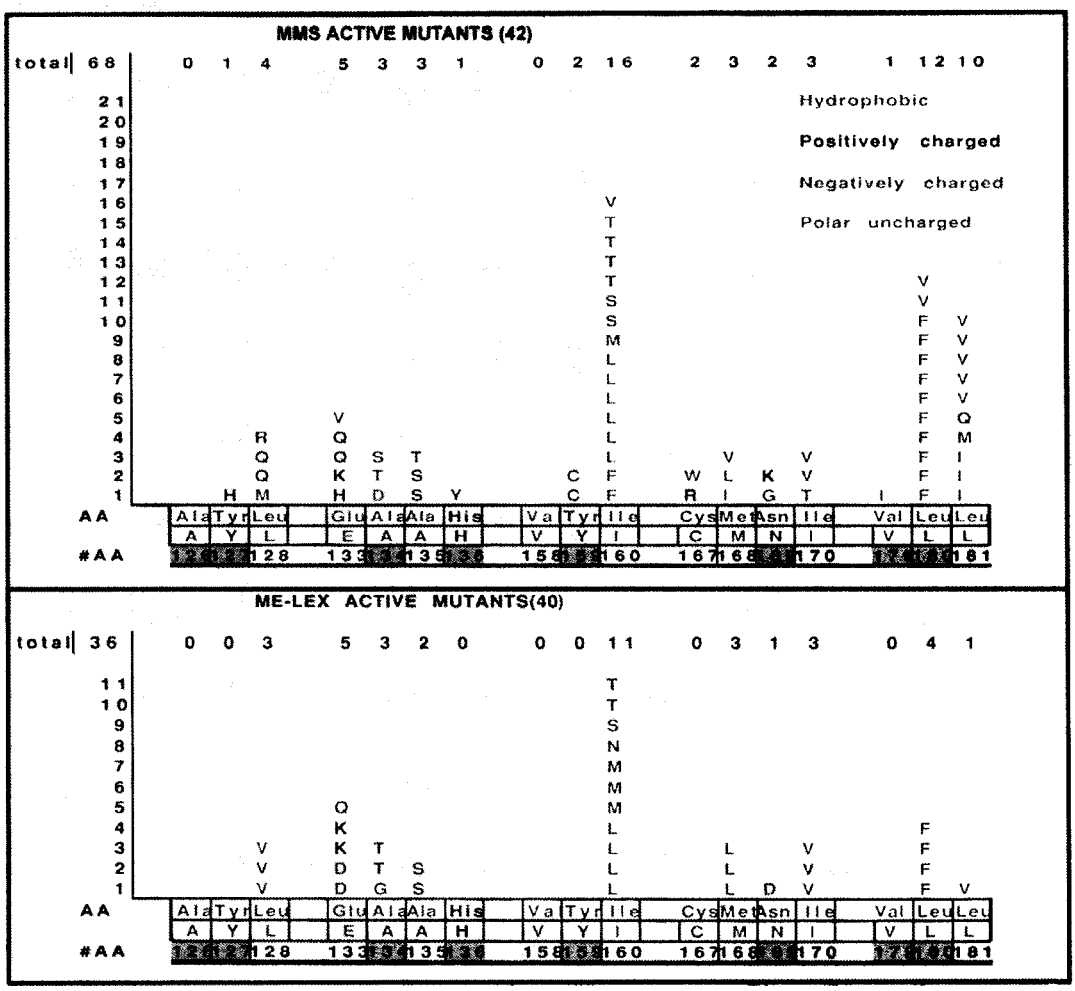



**Figure 24. Sequences of functional AAG mutants after selection.** The mutations were only made in 5 small regions between a.a. 126 and 181 instead of over the entire gene. 42 clones were used from MMS selection and 40 clones from Me-Lex selection. Any amino acid mutations in any of clones is shown in stacked columns. Below these rows, the wild-type amino acids for human AAG are shown above the corresponding amino acid numbers. Amino acids at the positions highlighted in blue are absolutely conserved throughout AAG homologues in human, mouse, rat, B. burgdorferi, B. subtilis, A. thaliana, and M. tuberculosis. Amino acids at positions highlighted in yellow are limited to two possible amino acid choices in these homologues. Note that there are some evolutionarily conserved residues that can tolerate limited mutations. The residues that are not evolutionarily conserved can tolerate an even larger number of substitutions.

Figure provided by Haiwei Guo and Larry Loeb.

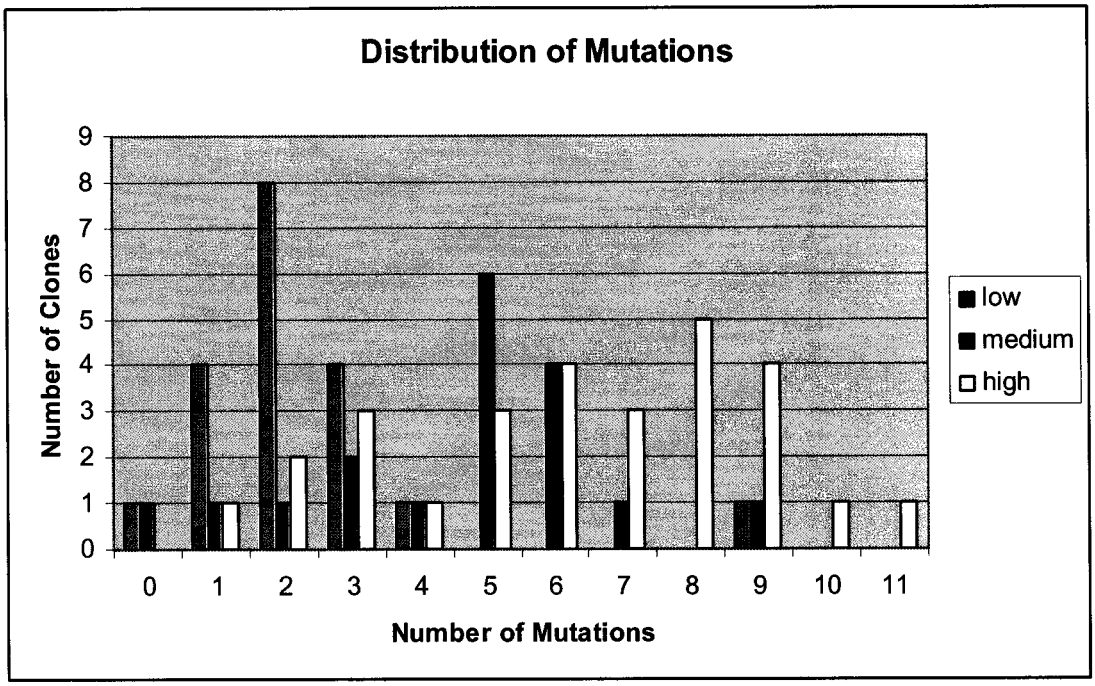



**Figure 25. Distribution of mutations in three mutant libraries.** The number of clones containing a given number of mutations is plotted. Frameshift errors are simply counted as one mutation in these calculations.

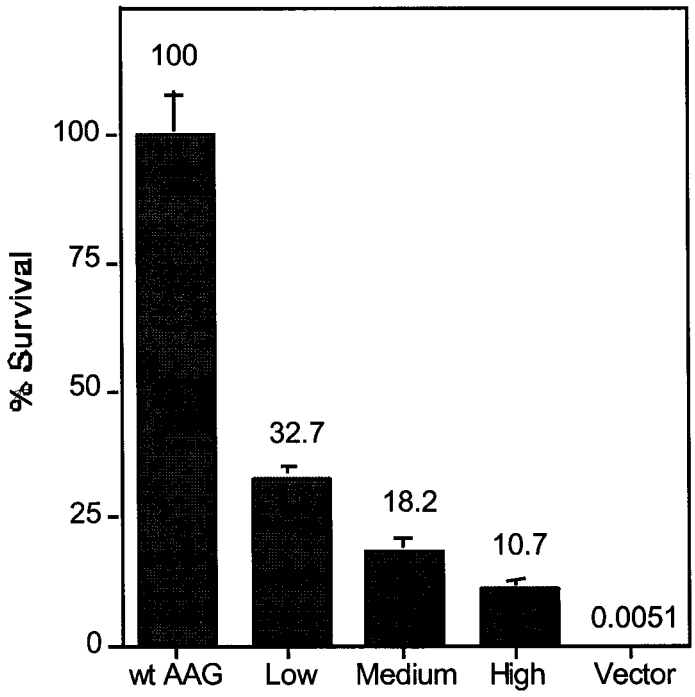



**Figure 26. Survival data for all three human AAG libraries.** The survival for wtAAG, and native pGRFP2 vector are also shown. Equivalent numbers of MV1932 cells transformed with each library were plated onto three normal LB plates and three LB plates containing 0.2% MMS. Percent survival was calculated by dividing the observed numbers of colonies on each plate. Note that wtAAG complements the MV1932 cells very well. With empty pGRFP2 there is almost no survival.

Figure provided by Haiwei Guo

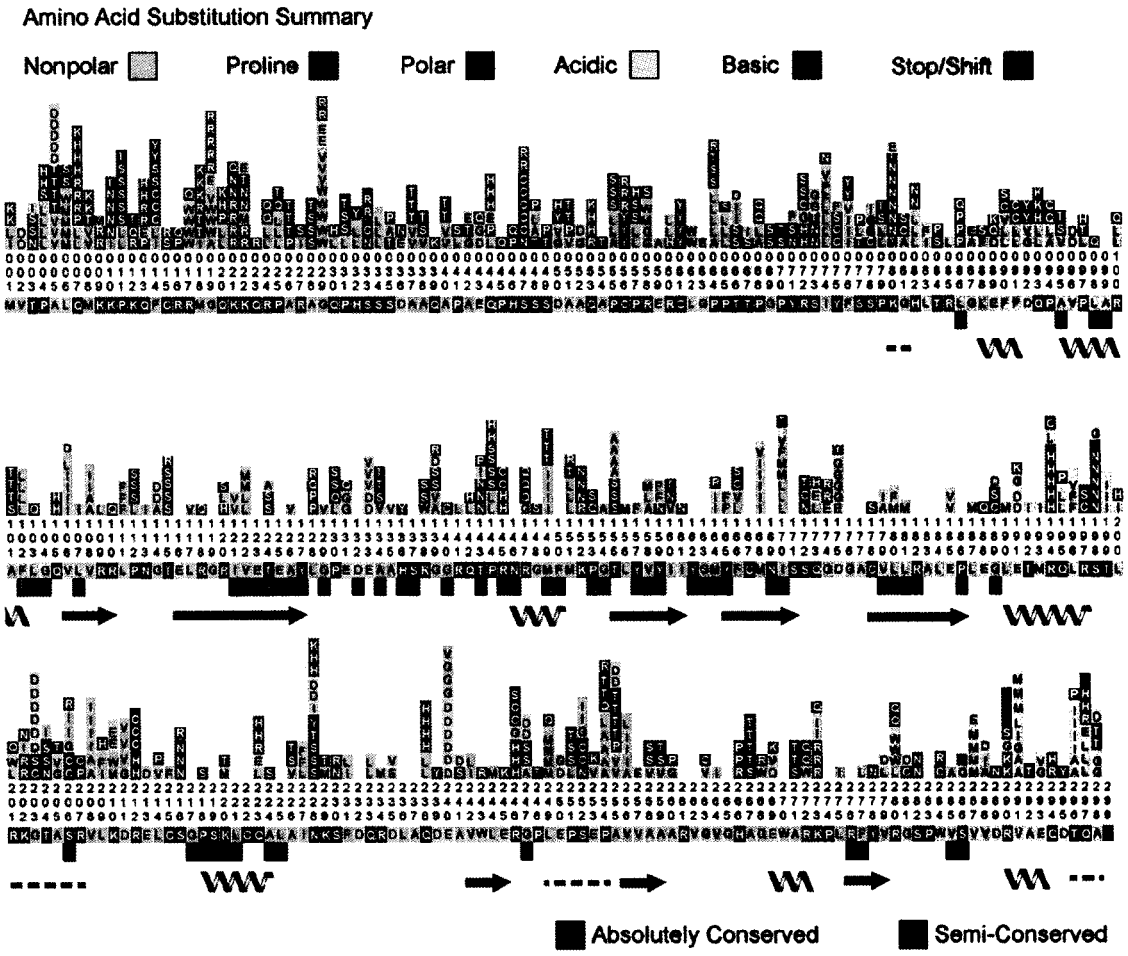



**Figure 27. Map of frequency, types, and locations of tolerated mutations in AAG.** The numbers and colored boxes below them indicate the amino acids in wild type AAG. Above the numbers, the height of each bar represents the total frequency at which mutations are found at that amino acid position. The colors represent various types of substituted amino acids. The letters in the colored bars represent observed mutations. The grey and black squares represent amino acids that are conserved evolutionarily according to Lau et al. Black boxes represent absolute conservation, and grey boxes represent partial conservation. Below the conservation markings, secondary structure elements are marked. The spirals indicate alpha helices, arrows indicate beta turns, and dashed lines indicate disordered loops.

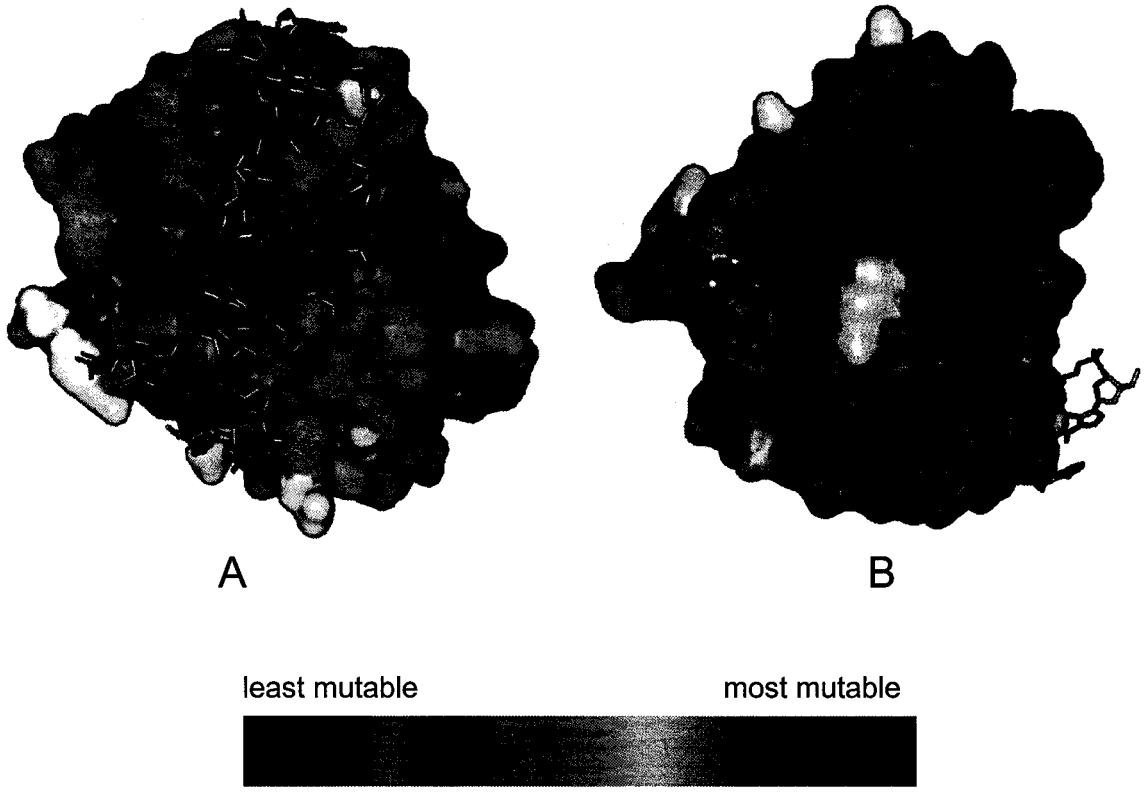



**Figure 28. 3D models of AAG with amino acids color-coded for mutability.** The mutability scores are based on the mutability scores calculated in this study and shown in figure 26. The purple and blue residues are least mutable and most sensitive to mutation. The yellow and green residues are slightly more mutable. The red residues are highly mutable. A) This figure shows the AAG protein bound to double stranded DNA. Notice that many of the regions that directly contact DNA are highly sensitive to mutation. B) This figure shows the back side of AAG facing away from the bound DNA. Many of the residues on the exterior have moderate tolerance to mutations without loss of function of the enzyme.

Figure produced by Haiwei Guo and Greg Ireton.

# *Chapter 4. Novel Method for Detecting DNA-Protein Interactions*

(In collaboration with Elijah Wallace and Ray Monnat, University of Washington, Seattle, WA)

## *A) Background*

With the sequencing of the human genome and with the accumulation of genomic sequences from countless other species, knowledge about the content and organization of genes is becoming readily available. The next step in our understanding of systems biology will come from decoding the "transcriptomes" and "proteomes" that are critical to understanding how genes work together to perform cellular activities. It is critical to know when genes are activated and in response to what stimuli. The process of transcriptional regulation is therefore an intensely studied subject.

Knowing when a gene is expressed often provides strong clues as to its biological function. This is the basis for microarray studies that study the mRNA levels of thousands of genes from cells in different states. The difference in mRNA levels between different cell populations indicates the selective activation or repression of genes.[1] Through microarray analysis, it has become apparent that many genes are activated or repressed at the same time in response to cellular states. These programs of gene expression can involve thousands of genes and are modified

[1] DeRisi, J.L., V.R. Iyer, and P.O. Brown. (1997). "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale." Science **278**: 680-686.

according to cellular states such as cell cycle, environmental stimuli, or during organismal development.[2]

These gene expression programs depend heavily on the function of transcriptional regulatory proteins that bind in the vicinity of promoter sequences. Transcription factor binding along with interactions with other component proteins modify the rate of transcription at that site.[3] It is gradually becoming clear that many of these cofactors are involved in alterations of chromatin structure to change the accessibility of DNA to transcriptional machinery.[4] Still other cofactors contact transcriptional machinery directly to alter the rate of transcription by RNA polymerase II (in eukaryotes).

Gene expression programs organize themselves into networks based on regulatory proteins and their corresponding binding sites. Some of the regulatory network motifs discovered by Lee et al. are shown in figure 29. As can be seen from this figure, regulatory networks can contain complex patterns. One can also imagine how simple regulatory network motifs can be combined into larger and more complex networks. The fact that greater than 5% of our genes are predicted to encode transcription factors[5] underscores the potential complexity of the task of decoding these networks.

---

[2] Lee, T.I., N.J. Rinaldi, F. Robert et al. (2002). "Transcriptional Regulatory Networks in Saccharomyces cerevisiae." Science **298**: 799-804.
[3] Orphanides, G. and D. Reinberg. (2002) "A Unified Theory of Gene Expression." Cell **108**: 439-451.
[4] Narlikar, G.J., H.-Y. Fan, and R.E. Kingston. (2002). "Cooperation between complexes that regulate chromatin structure and transcription." Cell **108**: 475–487.
[5] Tupler, R., G. Perini, and M.R. Green. (2001). "Expressing the human genome." Nature **409**: 832–833.

Critical to the task of decoding regulatory networks is the sensitive and specific detection of these protein-DNA binding events. Up until recently, there were very few methods to detect these protein-DNA interactions on a large scale. Ren et al. published a method within the past three years that is capable of identifying many of the target genes bound by a particular transcriptional regulator.[6] This method is sometimes referred to as genome-wide binding or location analysis. Figure 30 illustrates the process of Ren et al. Cells are treated with formaldehyde to crosslink any proteins to DNA they may be bound to. The cells are then lysed and the transcription factor that is being studies in immunoprecipitated with a specific antibody. The crosslinked DNA is coimmunoprecipitated in this process. The crosslinks are reversed, and the DNA is amplified and labeled with Cy5 fluorescent dye using ligation-mediated-PCR. An unenriched DNA pool is correspondingly labeled with Cy3. These two DNA pools can then be hybridized to a microarray containing intergenic sequences as determined from the genomic sequence of the organism being studied. An example of such a hybridization is shown in figure 31. One of the spots in this figure contains DNA that has been enriched through the immunoprecipitation process. This sequence contains a putative binding site for the transcription factor being studied.

The method of Ren et al. has been used in several studies to find genomic sites bound by transcription regulators [7,8,9] and DNA synthesis regulators[10] in yeast. Lee et

---

[6] Ren, B., F. Robert, J.J. Wyrick et al. (2000). "Genome-Wide Location and Function of DNA Binding Proteins." Science **290**:2306-2309.

[7] Iyer, V.R., C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, P.O. Brown. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**: 533-538.

al. applied this method on a large scale to determine the locations of binding sites of most of the transcriptional regulators encoded in *Saccharomyces cerevisiae.*[11] This method has been shown to work well in a large number of situations.

As with all methods, there are also some minor shortcomings to the method of Ren et al. Typically, a false positive rate of 6-10% is observed even in the highest quality data.[12] Therefore, further studies must be conducted to confirm protein-DNA interactions.

Also, this method does not work with many transcription factors, because the protein is not found in large quantities inside cells. Protein-DNA complexes cannot be immunoprecipitated efficiently if there are very few copies of the protein present.

Another complication is that antibodies need to be generated for every transcription factor tested. Alternatively, strains of cells can be engineered to contain a homogeneous tag fused to the transcription factor of interest.[13] Then, one antibody against the tag can be used for all transcription factors. Besides being time consuming, this adds some complexity since it is not known whether the tag will affect the DNA binding activity of a particular transcription factor.

---

[8] Lieb, J.D., X. Liu, D. Botstein, P.O. Brown. "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." Nature Genetics **28**: 327-334.
[9] Simon, I., J. Barnett J, N. Hannett N et al. (2001). "Serial regulation of transcriptional regulators in the yeast cell cycle." Cell **106**: 697-708.
[10] Wyrick, J.J., J.G. Aparicio, T. Chen et al. (2001). "Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins." Science **294**: 2357-2360.
[11] Lee, T.I., N.J. Rinaldi, F. Robert et al. (2002). "Transcriptional Regulatory Networks in Saccharomyces cerevisiae." Science **298**: 799-804.
[12] http://web.wi.mit.edu/young/regulator_network
[13] Lee, T.I., N.J. Rinaldi, F. Robert et al. (2002). "Transcriptional Regulatory Networks in Saccharomyces cerevisiae." Science **298**: 799-804.

Because the pieces of DNA isolated through the immunoprecipitation process are fairly large, the protein binding site can only be identified to within several kilobytes. If there are regulatory regions for several genes within this DNA fragment, then further studies must be conducted to resolve this ambiguity.

Finally, the Ren et al. method relies on microarray technology to identify the protein binding regions brought down by immunoprecipitation. In designing the microarray, a great deal of information is required about the genomics of the species being studied. Besides having the genomic sequence, putative genes and intergenic regions must be identified. Virtually all of the studies performed to date have been on yeast in which the genome is relatively small. A great deal is known about the locations of all the yeast genes. When scaling up to more complex eukaryotic species, one major limitation will be the availability of limited genomic information. Many species in which investigators might desire to study transcription factors do not have a completed genomic sequence or have limited information on the locations of genes.

Due to some of these shortcomings, a second complimentary method for identifying DNA binding sites for DNA binding proteins could be very useful to investigators. In this work, we present proof of concept of such a method using a modified form of the pGRFP series of vectors.

### *B) Vector Design and Basis for DNA-Protein Binding*

In this section, we describe the design of the pGFPpDsRed series of vectors that can be used to identify protein-DNA interactions within the intracellular environment

of *E. coli* cells. The basic vector design is shown in figure 32A and 32B. The GFP and DsRed genes are now expressed from two identical but separate promoters. Read through transcription is prevented with a transcriptional terminator. The putative binding site is cloned into the vector so that it is transcribed upstream of the DsRed gene and its ribosomal binding site (RBS). The putative DNA binding protein is expressed from a separate expression plasmid that is co-transformed into *E. coli* along with the pGFPpDsRed series vector. An example of such an expression plasmid is shown in figure 32C.

If there is no binding of the expressed DNA binding protein to its binding site, then there will be approximately equal transcription of GFP and DsRed. This is because they are being expressed from the same plasmid from identical promoters. If there is binding of the DNA binding protein to the cloned binding site, then it is proposed that transcription of DsRed will become slightly more inefficient as compared with GFP. The presence of a bound protein on double stranded DNA would delay the transcriptional machinery of *E. coli* from reading through the region of the binding site. Decreased DsRed transcription would manifest itself as decreased red fluorescence in *E. coli* cells as compared with green fluorescence from GFP. In essence, the constitutive expression of GFP serves as a control by which to interpret the level of expression of DsRed.

If a library of possible DNA binding sites is cloned into pGFPpDsRed, then we should be able to distinguish clones containing protein binding sites based on the ratio of red to green fluorescence of each cell. These cells can be rapidly flow sorted and

cultured. At this point, multiple rounds of culturing and flow sorting can be performed to further enrich for binding sites. The remaining clones can then be clonally isolated and sequenced, similar to the process described in the "Sequencing Project" chapter. If the library of potential DNA binding sites is created from small pieces of genomic DNA, then the entire genome of a species could be quickly scanned to look for binding sites for transcription factors or any other DNA binding proteins. This full process is shown in figure 33. If the pieces of genomic DNA are approximately 100bp in size, then these binding sites will be localized to a very small region of the genome. It should be relatively easy to determine which genes are being controlled by a particular protein. The actual genomic scan should proceed very quickly. If a flow cytometer is examining a 100bp genomic library at a rate of 30,000 cells per second, we could scan the yeast genome to 1X coverage every 4 seconds. For the human genome, we would obtain 1X coverage every 17 minutes. We could obtain a very dense scan of human genomic DNA for binding sites in the course of a single day.

Alternatively, the binding site cloned into pGFPpDsRed could be kept constant, and a library of potential binding proteins could be cloned into the expression plasmid. This would be especially useful for mutagenesis studies in which investigators want to generate new DNA binding proteins with altered binding specificities. A slight alteration could be made in the original binding sequence, and the altered version could be cloned into the pGFPpDsRed vector. Then, a mutant library of the original DNA binding protein could be created by PCR or

oligonucleotide insertion mutagenesis. The flow cytometer could then be used to quickly isolate clones with successful DNA-protein binding. The mutated protein sequence could then be sequenced to determine the alterations in amino acid sequence.

There is good evidence to suggest that bound proteins would be able to effectively hinder or even halt the process of transcription. Repressor proteins have evolved specifically to perform this vary task. The mechanism of action of many different repressor proteins has been well studied over the years. There is a widely accepted view that repressor proteins bound to DNA simply deny RNA polymerases access to the necessary promoter regions.[14,15] Lutz, et al. quantitatively showed that the lac repressor significantly slows the kinetics of RNA polymerase binding.[16] There is some emerging evidence, however, that RNA polymerases can still bind to the promoter in the presence of some repressors. Blocking of transcription has been shown to occur during open complex formation[17] or promoter clearance[18] with the lac repressor.

---

[14] Schlax, P.J., M.W. Capp, and M.T. Record. (1995). "Inhibition of transcription initiation by lac repressor." Nucl. Acids Res. **6**:111-137.

[15] Bertrand-Burggraf, E., S. Hurstel, M. Duane, and M. Schnarr. (1987). "Promoter properties and negative regulation of the uvrA gene by the LexA repressor and its amino-terminal DNA binding domain." J. Mol. Biol. **193**: 293-302.

[16] Lutz, R., T. Lozinski, T. Ellinger, and H. Bujard. "Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator." Nuc. Acids Res. **29**:3873-3881.

[17] Straney, S.B. and D.M. Crothers. (1987). "lac repressor is a transient gene-activating protein." Cell **51**:699-707.

[18] Lee, J. and A. Goldfarb. (1991). "*lac* repressor acts by modifying the initial transcribing complex so that it cannot leave the promoter." Cell **66**:793-798.

Lopez et al. studied the mechanisms by which lac repressor inhibits transcription by the T7 RNA polymerase from a T7 promoter.[19] They show that T7 RNA polymerase is able to bind and initiate transcription. However, if the lac repressor is bound to the double stranded DNA just downstream from the promoter, then the transcription complex falls off after transcribing only a few bases. They hypothesize that this occurs, because the RNA polymerase has a high disassociation rate (low processivity) during the early stages of transcription.[20] Because the polymerase has a lower binding affinity than the lac repressor protein to the lac operator site, the polymerase usually falls off before the repressor protein does. This results in aborted transcription. Further from the promoter, the processivity of the transcription complex increases. Therefore, bound repressor protein has less inhibitory effect on transcription.[21]

Lopez et al. made a set of promoter constructs with the lac operator site inserted at various positions relative to the start of transcription.[22] When the center of the lac operator was situated at the +13 or +15 positions, there was almost complete inhibition of transcription with almost no detectable full length transcript. Many aborted transcription products only a few ribonucleotides long were seen. In the construct with lac operator placed at +47, there was much less inhibition of read-through transcription. However, it is interesting to note that full length transcript was

---

[19] Lopez, P.J., J. Guillerez, R. Sousa, and M. Dreyfus. (1998). "On the Mechanism of Inhibition of Phage T7 RNA Polymerase by lac Repressor." J. Mol. Biol. **276**:861-875.
[20] Martin, C.T., D.K. Muller, and J.E. Coleman. (1988). "Processivity in early stages of transcription by T7 RNA polymerase." Biochemistry **27**: 3966-3974.
[21] Lopez, P.J., J. Guillerez, R. Sousa, and M. Dreyfus. (1998). "On the Mechanism of Inhibition of Phage T7 RNA Polymerase by lac Repressor." J. Mol. Biol. **276**:861-875.
[22] Ibid.

still diminished by approximately ½, even after the T7 polymerase has reached full processivity. This indicates that polymerases may be most sensitive to repression when the bound protein is located close to the promoter. However, detectable repression still occurs when protein is bound to double stranded DNA at points further downstream from the promoter. The degree of repression will depend on the processivity of the RNA polymerase compared with the affinity for the bound protein to its DNA binding site.

In our vector system, the multiple cloning site (MCS) is located fairly close to the promoter. The first restriction site, BamHI, is located only 4 bases away from the start of transcription point. If a protein binding site is cloned into the MCS, it has the potential to be situated very close to the promoter for maximum inhibition of transcription. However, the Lopez et al. data show that we may get appreciable repression of transcription even if the binding site is located further from the promoter. This may occur if a large fragment is cloned into the pGFPpDsRed vector and the binding site is located fairly far downstream. Also, if the insertion fragments are randomly sheared pieces of genomic DNA, protein binding sites will be situated with random distance from the promoter. Because we have the capability to rapidly oversample genomic DNA to high redundancy, we will likely see clones in which any given protein binding site embedded in genomic DNA will be located very close to the promoter. These clones will likely display significantly repressed transcription of DsRed when a protein is bound to the cloned site. This will result in a high ratio of

GFP fluorescence to DsRed fluorescence, and these clones can easily be isolated by flow cytometry.

### *C) Vector Construction*

The pGFPpDsRed series of vectors are derived directly from the pGRFP series of vectors. The new pGFPpDsRed2 vector contains a stop codon and an rrnB transcriptional terminator after GFP. This results in the constitutive expression of the GFP gene under the control of its lac promoter. Transcription then ends shortly afterwards at the transcriptional terminator. A new promoter region has been inserted in front of the DsRed-T3 gene. This region contains an identical lac promoter to the promoter used for GFP. However, after the start of transcription point, there is a multiple cloning site (MCS) that can be used to clone in candidate protein binding sites. A 50 bp linker region separates the protein binding site from a consensus RBS and the start codon of the DsRed-T3 gene.

The pGFPpDsRed series vectors were constructed in multiple stages. The initial pGFPpDsRed vector was produced from the pGRFP vector (see chapter entitled "Sequencing Project"). The pGRFP vector was cut at the NotI and SalI sites between the GFP and DsRed genes, and the resulting linear fragment was 5' dephosphorylated and gel purified. We constructed a 576bp cassette containing (in order) a stop codon, rrnB transcriptional terminator, lac promoter, MCS, RBS, and new start codon for insertion into pGRFP. The stop codon would terminate translation of the GFP gene. The rrnB transcriptional terminator would prevent read-through transcription from the

GFP gene into the DsRed gene. The new lac promoter is identical to the promoter in front of GFP and will initiate transcription of the DsRed gene separately from GFP. The MCS is located after the start of transcription start point and is the site where candidate protein binding sites will be cloned. This is a 41bp stretch with the following unique restriction sites: BamHI, BsaBI, BspDI, ClaI, RcoRV, NotI, EagI, NspI, SphI, SapI, and SalI. The consensus RBS and start codon follow the MCS and are necessary to initiate translation of the DsRed gene. They are identical to the sequences in front of the GFP gene.

The majority of the 576bp cassette including the stop codon and rrnB transcriptional terminator was PCR amplified from a pSE280 plasmid (Invitrogen) using Pfu polymerase (Stratagene). The final PCR product length was ~450bp in length. The upstream primer added a PspOMI restriction site in order to later interface to the NotI site. The lac promoter, MCS, RBS, start codon, and a XhoI site were produced from two oligonucleotides 116bp and 111bp in length. They were designed to overlap 23 bases. The two oligonucleotides were placed in PCR reaction buffer and thermocycled with Pfu polymerase to extend from the 3' ends of the oligonucleotides. This yielded an ~200bp double stranded DNA product. This product was designed to overlap 33 bases with the PCR fragment from pSE280. Due to the overlap, we were able to "ligate" these two pieces together by adding both products to a PCR reaction and amplifying using two primers, one that annealed to the upstream portion of the ~450bp PCR product and one that annealed in reverse to the downstream portion of the ~200bp extension product. This yielded an ~600bp product. This product was

digested with PspOMI and XhoI to create sticky ends. Some leader sequence on the ends of the PCR product allowed efficient cutting. The PspOMI/XhoI ends were directly ligated into the NotI/SalI ends of the pGRFP2 fragment. We called this vector pGFPpDsRed.

The initial pGFPpDsRed vector was sequence verified, but we could not observe any red fluorescence from colonies or cells. Bright GFP fluorescence was still present. DsRed was not detectable by Western blot either. We entered the predicted mRNA sequence into the online MFOLD program[23] and discovered that thermodynamically stable hairpins were forming near the vicinity of the RBS. By experimenting with different designs, we discovered that hairpins were commonly forming between sequence in the MCS and sequence after the RBS. By obscuring the RBS and start codon, translation of DsRed will be severely limited.

Therefore, we designed a second generation vector. This new vector would contain the new DsRed-T3 variant that worked so well in the pGRFP2 vector (see the "Sequencing Project" chapter). Also, the early transcribed region was redesigned with a slightly modified MCS, a specially designed inert 50bp linker between the MCS and RBS, and silent mutations in codons 2,3, and 4 of the DsRed-T3 gene. The new MCS contains the following unique restriction sites: BamHI, BsaBI, BspDI, ClaI, RcoRV, NotI, EagI, NspI, SphI, SalI. All of these changes were designed to prevent any

[23] Zuker, M., D.H. Mathews and D.H. Turner. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology. Kluwer Academic Publishers, 1999.

energetically favorable secondary structures from forming that might inhibit translation.

We PCR amplified the DsRed-T3 gene from the pGRFP2 vector with phosphorylated primers, The forward primer had additional bases at its 5' portion consisting of a seven base spacer, SalI restriction site, a 50bp inert linker sequence after the MCS, the RBS, start codon, and replacement for codons 2,3, and 4 of the gene. The 3' portion of the forward primer annealed to the DsRed-T3 gene starting from codon 5. This yielded a 784bp product. The majority of the pGFPpDsRed vector was PCR amplified starting from after the wtDsRed gene, through the origin of replication, through the GFP gene, rrnB transcriptional terminator, new DsRed promoter, and most of the MCS up through the SphI site. This yielded a 3226bp PCR product. Because the two pieces were PCR amplified using Pfu polymerase, the end products had blunt ends. The two pieces were blunt-end ligated together and clones were sequence verified. This new vector was called pGFPpDsRed2.

In early testing, pGFPpDsRed2 seemed to display bright GFP and DsRed fluorescence. The translation problems inherent in the pGFPpDsRed design seemed to have been resolved. We also designed a third vector, pGFPpDsRed3 that replaced the pUC origin with a lower copy number pBR322 origin. We PCR amplified a 2189bp region from pET15b (Novagen) containing the rop protein, pBR322 origin of replication, and part of the AmpR gene. We PCR amplified a 2609bp region from pGFPpDsRed2 using phosphorylated primers. The amplified region contained the GFP and DsRed genes with associated promoter regions and part of the AmpR gene.

We blunt-end ligated these two PCR products. A functional AmpR gene would only be reconstructed if the two PCR products were ligated in the correct orientation. A clone with the correct sequence was chosen after restriction analysis.

Both the pGFPpDsRed2 and pGFPpDsRed3 vectors proved somewhat difficult to work with. Because we are trying to balance the transcription from the GFP and DsRed promoters, subtle changes can sometimes lead to discrepancies in the fluorescences from these two proteins. There was a very large variation in fluorescence ratio between cells when using certain strains of *E. coli.* This is confirmed by the work of Elowitz et al. who expressed CFP and YFP in bacterial cells from identical promoters.[24] They noted that "intrinsic noise" in bacterial systems often causes random discrepancies in the expression of proteins, even when identical promoters are used. They found large differences in certain bacterial strains in the level of "intrinsic noise". In particular, strains without RecA have higher "intrinsic noice".

In our hands, transforming DH10B cells (Invitrogen) with either pGFPpDsRed2 or pGFPpDsRed3 resulted in the phenotypic segregation of cells into a predominantly green and predominantly red fluorescing populations. When these populations were sorted and grown, they yielded identical appearing cultures again. This indicates that the cells are segregating into green or red phenotypes independent of genetic changes. It is not clear exactly why this occurs. When XL1-Blue cells (Strategene) are grown under identical conditions, this phenotypic segregation does not occur, and

---

[24] Elowitz, M.B., A.J. Levine, E.D. Siggia, and P.S. Swain. (2002). "Stochastic Gene Expression in a Single Cell." Science **297**: 1183-1186.

fluorescence levels are much higher. Both of these strains used are RecA minus in order to reduce the frequency of deleterious rearrangements.

We typically grew cultures at 37 degrees celcius for approximately 12 hours under 0.5% glucose repression. After cultures reached late log or stationary phase, the glucose was removed and cells were incubated for up to 24 hours. During this time, the cells produce fluorescent proteins which then mature to their fluorescent state. The final incubation temperature seems to be very important in determining how homogeneous cells' GFP to DsRed fluorescence ratios are. For pGFPpDsRed3, cells incubated at 37 degrees formed a much tighter population compared with cells incubated at 30 degrees.

Overall, the copy number of pGFPpDsRed3 vector is about 10 times less than the copy number of pGFPpDsRed2. This translates into significantly decreased measured fluorescence of cells. However, the pGFPpDsRed3 vector seems to also yield more consistent results. This lower rate of protein synthesis may have something to do with this consistency, since cells may be under less stress. Also, if there are less copies of plasmid, it may be easier for any expressed DNA-binding proteins to saturate the majority of binding sites. Therefore, we primarily used pGFPpDsRed3 for our experiments.

### *D) Initial Results*

Our goal in this work is to show proof of concept of the pGFPpDsRed series of vectors as viable tools for detecting DNA-protein interactions. We used the

interaction of STE12 yeast protein with its corresponding binding site in these experiments. The STE12 gene product is a well studied transcription factor involved in the pheromone response pathway of *Saccharomyces cerevisiae*. The binding site with sequence "ATGAAACA" has been shown to be necessary for induction of certain genes in the pheromone response pathway.[25] It has also been shown that the STE12 protein binds specifically to the "ATGAAACA" sequence in the context of the yeast genome as well as when a double copy is placed into a pUC18 bacterial plasmid.[26]

We constructed a double stranded DNA containing the STE12 binding site using two annealed 5' phosphorylated 20bp oligonucleotides. A 5' four base overhang on both sides makes the fragment compatible with BamHI ends. The STE12 binding site fragment is shown here:

```
5'-GATCCCATATGAAACAAATG      -3'
3'-     GGTATACTTTGTTTACCTAG -5'
```

We incubated this fragment by itself in a ligration reaction for approximately 2 hours. Then, we added BamHI cut 5' dephosphorylated pGFPpDsRed3 vector at a 8:1 insert to vector ratio. After transforming and plating cells, we sequenced a number of colonies. We selected three colonies for further study: one contained the STE12 binding site in the reverse orientation (R), one contained two STE12 binding

---

[25] Kronstad, J.W., J.A. Holly, and V.L. MacKay. (1987). "A yeast operator overlaps an upstream activation site." Cell **50**: 369-377.

[26] Dolan, J.W., C. Kirkman, and S. Fields. (1989). "The yeast STE12 protein binds to the DNA sequence mediating pheromone induction." Biochemistry **86**: 5703-5707.

sites in the forward-reverse orientations (FR), and one contained three STE12 binding sites in the forward-reverse-reverse orientations (FRR).

We made an expression plasmid, pACYC-STE12, to produce STE12 protein. This plasmid is shown in figure 32C. The pACYC vector backbone was used, because it is a compatible plasmid with the pBR322 based vectors used for fluorescent protein expression. These two plasmids can therefore co-exist within the same *E. coli* cells. We PCR amplified the vector backbone from the standard plasmid, pACYC184. A second PCR product was produced from a pSE380 plasmid (Invitrogen) containing a lacIq gene, Trc promoter, and part of a multiple cloning site. These two PCR fragments were blunt-end ligated together and the resulting product, pACYC-Trc, was verified by restriction mapping. The STE12 gene product was amplified as part of a pre-constructed GST-STE12 protein fusion from a pGEX series plasmid (generous gift from John Aitchison, Institute for Systems Biology, Seattle, WA). The forward primer contained a consensus ribosomal binding site that was added to the fragment a short distance from the start codon of GST. This PCR fragment was blunt-end cloned into the BmgBI site of the pACYC-Trc vector. Clones were confirmed by sequence analysis.

We co-transformed the pACYC-STE12 plasmid with either native pGFPpDsRed3 with no STE12 binding sites or pGFPpDsRed3 with variable numbers of STE12 binding sites (R, FR, FRR). Cells were grown overnight on LB plates containing both ampicillin and chloramphenicol to select for cells containing both plasmids. Colonies were picked into 5 ml. LB media containing 50μg/ml

carbenicillin, 12.5μg/ml chloramphenicol, and 0.5% glucose and grown at 37 degrees shaken for 12 hours. The glucose represses fluorescent protein production. The cultures were pelleted and resuspended in 5 ml. LB media containing 50μg/ml carbenicillin, 12.5μg/ml chloramphenicol, and 1mM IPTG. This medium allows production of the STE12 protein as well as the two fluorescent proteins. The cultures were incubated for 24 more hours at 37 degrees shaken.

The cultures were analyzed by flow cytometry with identical setup as the experiments with the pGRFP series vectors (see "Sequencing Project" chapter). The dot plots showing the GFP and DsRed fluorescence values are shown in figure 34. Figure 34A shows the culture from native pGFPpDsRed3. Figures 34B, C, and D. show the cultures from pGFPpDsRed3 with one, two, and three STE12 binding sites (in the "R", "FR", and "FRR" orientations) respectively. There is a very distinctive diminishment in the fluorescence of DsRed with one STE12 site. The cultures with two and three STE12 sites have even more remarkably diminished DsRed fluorescence in a dose-dependent response. This dose-dependent response seems to suggest that we are actually seeing repression of DsRed transcription due to STE12 protein binding rather than seeing some random artifacts caused by culture conditions.

We then mixed equal volumes of cultures to create approximately 50/50 mixes of pGFPpDsRed3 containing 0 and 1 STE12 binding sites and pGFPpDsRed3 containing 0 and 3 STE12 binding sites. We ran these mixed cultures in the flow cytometer and obtained the dot plots shown in figure 35. There is a slight separation

in GFP:DsRed fluorescence ratios between the populations containing 0 and 1 STE12 binding sites. There is an extremely large separation between populations containing 0 and 3 STE12 binding sites.

We flow sorted from these mixed cultures selecting the cells with the lowest DsRed fluorescence relative to GFP fluorescence. The cells within the sort gate are shown in green in figure 35A,B. These sorted cells constituted 2.8 and 3.4 % of total cells in the mixed 0 and 1 binding site pool and mixed 0 and 3 binding site pool respectively. Many cells were sorted directly onto agar plates and grown overnight. Nine colonies were chosen at random from each plate and cultured. Elijah Wallace (University of Washington, Seattle, WA) prepared plasmid DNA from these cultures and digested with Nde I to determine whether the sorted clones contained STE12 binding sites or not. Figure 36 shows the gel with the NdeI digests. NdeI normally cuts pGFPpDsRed3 at one site, yielding a 4.8kb fragment. However, the STE12 binding site cassette that we used also contains one NdeI site. If there are more than one STE12 site, a few small fragments will result that typically will not be visible by gel. The primary restriction fragments will be 3.8kb and 1.0kb. If there is only partial cutting of the plasmid, a 4.8kb fragment may also be visible that represents the linearized plasmid. Figure 36 shows that we are able to successfully differentiate between cells containing 0 and 1 STE12 binding sites per cell. Eight out of nine clones contained one STE12 binding site in the final analysis. In differentiating between cells containing 0 and 3 STE12 binding sites, we were able to select cells

with a high degree of accuracy. Nine out of nine of the sorted clones contained three STE12 binding sites.

These data show clear proof of concept that the pGFPpDsRed series of vectors is able to successfully detect DNA-protein interactions. From a mixed pool of cells with and without STE12 binding sites, we were able to select cells containing binding sites with a high degree of precision. This was even true when only one site was present. This means that the pGFPpDsRed system has the sensitivity to detect DNA-protein interactions at the level of a single bound protein per plasmid copy.

Clearly, more work is necessary to fully validate this system for application in the real world. There were several variables that were well controlled in this experiment to maximize the likelihood of success. One such factor is the distance of the cloned STE 12 binding site to the start of transcription point. At its closest the STE12 binding site started approximately 14bp into the transcript. This may be close enough where the disassociation constant for the RNA polymerase may still be high. As you move protein binding sites further away from the start of transcription point, we may observe less repression of DsRed transcription. For instance, it would be interesting to see if a STE12 binding site located 50bp away from the transcription start point would still measurably repress transcription. There is some weak evidence to suggest that repression would still occur at further distances. In the case of the pGFPpDsRed3 vector with three STE12 binding sites, the most distant STE12 binding site occurs 53 bases away from the start of transcription. However, this most

distant site differentially reduces DsRed transcription relative to the case in which there are only two STE12 binding sites (see figure 34).

In these experiments, we had a pre-designated STE12 binding site that we engineered into the pGFPpDsRed3 plasmid. In sorting cells from the mixed pools, the sorter only had to differentiate between clones with and without STE12 binding sites. It would be very informative to clone a complex genomic library into pGFPpDsRed3. In this case, there would literally be hundreds of thousands of distinct clones to choose from. This might complicate the picture in unforeseen ways. For instance, it is possible that some genomic elements might artificially repress DsRed transcription by mechanisms not involving protein binding. The goal of this experiment would be to try and sort cells with relatively low DsRed fluorescence compared with GFP fluorescence. If we see a significant enrichment for clones containing protein binding sites, then this will serve as further validation for this system of detecting DNA-protein interactions.

Using *E. coli* for this system has many advantages and disadvantages. It is advantageous to express eukaryotic DNA binding proteins in *E. coli*, because it is a relatively clean environment. In other words, the protein and DNA elements are interacting in an organism with completely different sets of proteins. It is relatively unlikely that many bacterial proteins would bind to eukaryotic binding sites or prevent the eukaryotic protein from binding. One major disadvantage of using *E. coli* is that not all eukaryotic proteins can be produced in functional form in *E. coli*. There are a large range of issues that could potentially interfere with a given protein's

function including a lack of post-translational modifications, solubility or other cytoplasmic issues, and the potential need for accessory proteins for binding. It may be possible to modify the coding sequence for the binding protein to solve these problems or to co-express accessory proteins. Alternatively, it may be possible to create analogous systems involving fluorescent proteins in yeast to solve some of these issues.

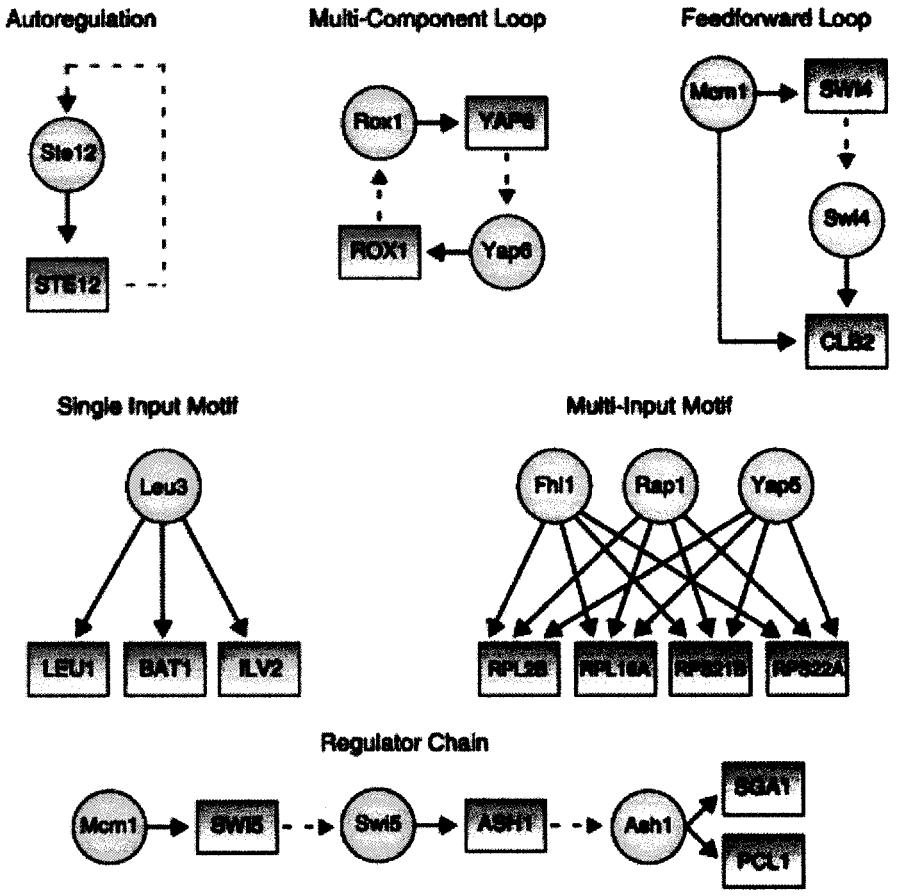
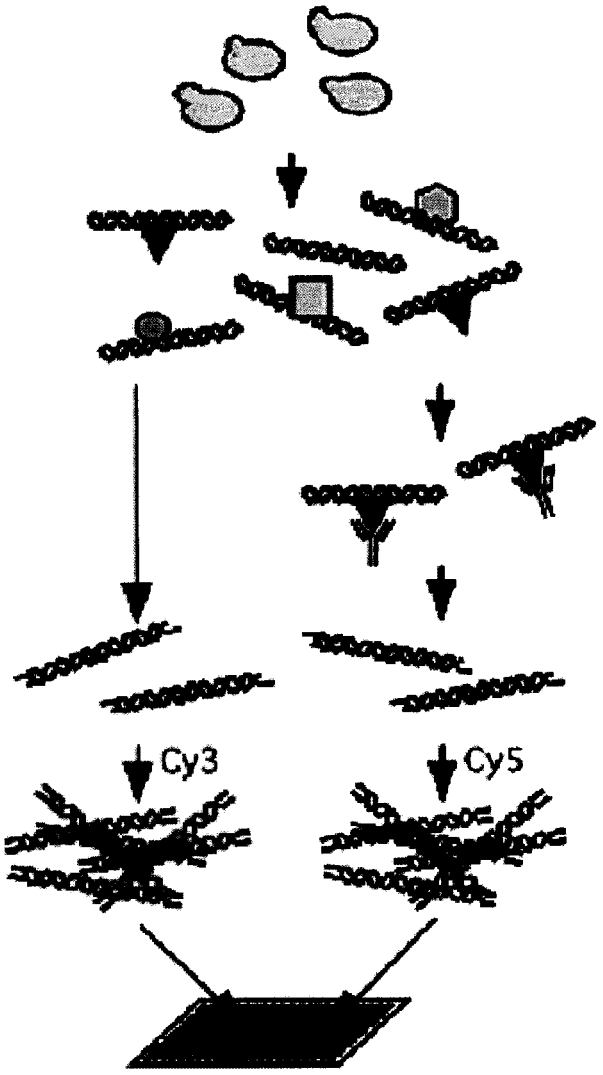



**Figure 29. Gene regulatory network motifs.** This figure from Lee et al. illustrates various motifs discovered by looking for patterns in their microarray data. The blue circles represent regulatory proteins. Red rectangles represent promoters and genes. The solid arrows indicate binding of regulatory proteins to protein binding sites. Dashed arrows represent the process of transcription and translation to produce a protein product from its encoding gene.

Figure reproduced from Lee et al. 2002.




Crosslink protein to DNA in vivo with formaldehyde

Break open cells and shear DNA

Immunoprecipitate

Reverse-crosslinks, blunt DNA and ligate to unidirectional linkers, LM-PCR

Hybridize to array

**Figure 30. Genome wide binding analysis method of Ren et al.** Cells are treated with formaldehyde to crosslink proteins and DNA they may be bound to. The cells are then lysed and the transcription factor that is being studies in immunoprecipitated with a specific antibody. The crosslinks are reversed, and the DNA is amplified and labeled with fluorescent dye. An unenriched DNA pool is also prepared with a different fluorescent dye. These two DNA pools can then be hybridized to a microarray containing intergenic sequences to find protein binding sites.

Figure reproduced from http://web.wi.mit.edu/young/location/

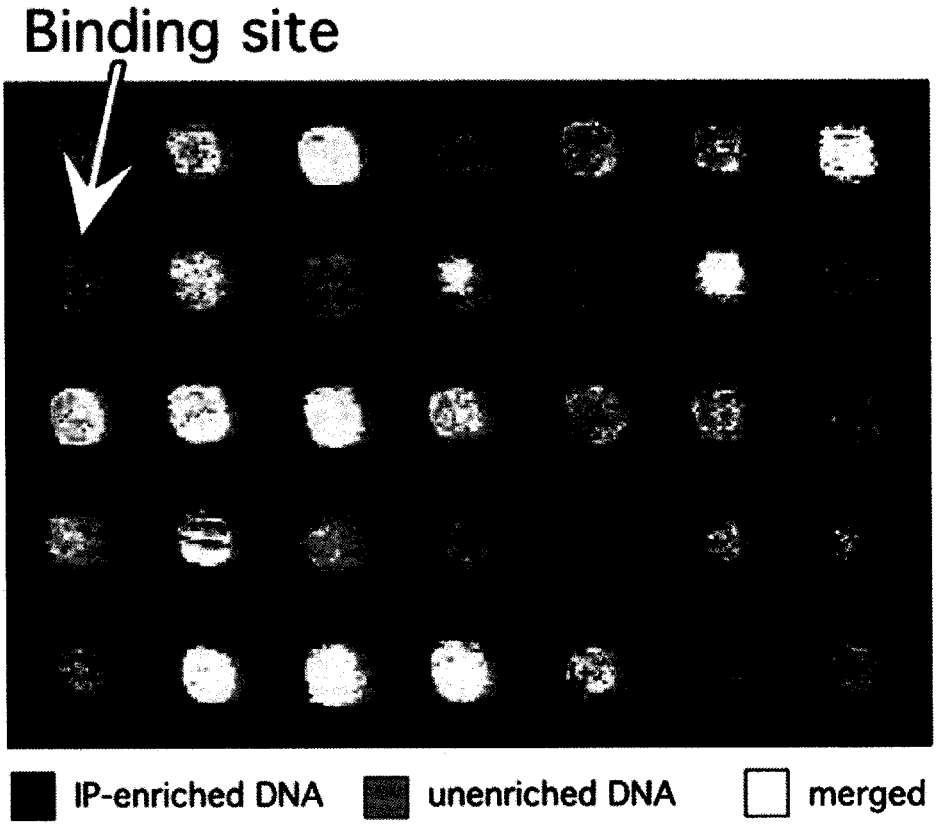



**Figure 31. Closeup of microarray from Ren et al. showing binding site.** This microarray was produced with yeast intergenic sequences. The red spot indicates a sequence that was enriched through the immunoprecipitation of a transcription factor. This sequence contains a putative binding site for the transcription factor being studied.

Figure reproduced from Ren et al. 2000.

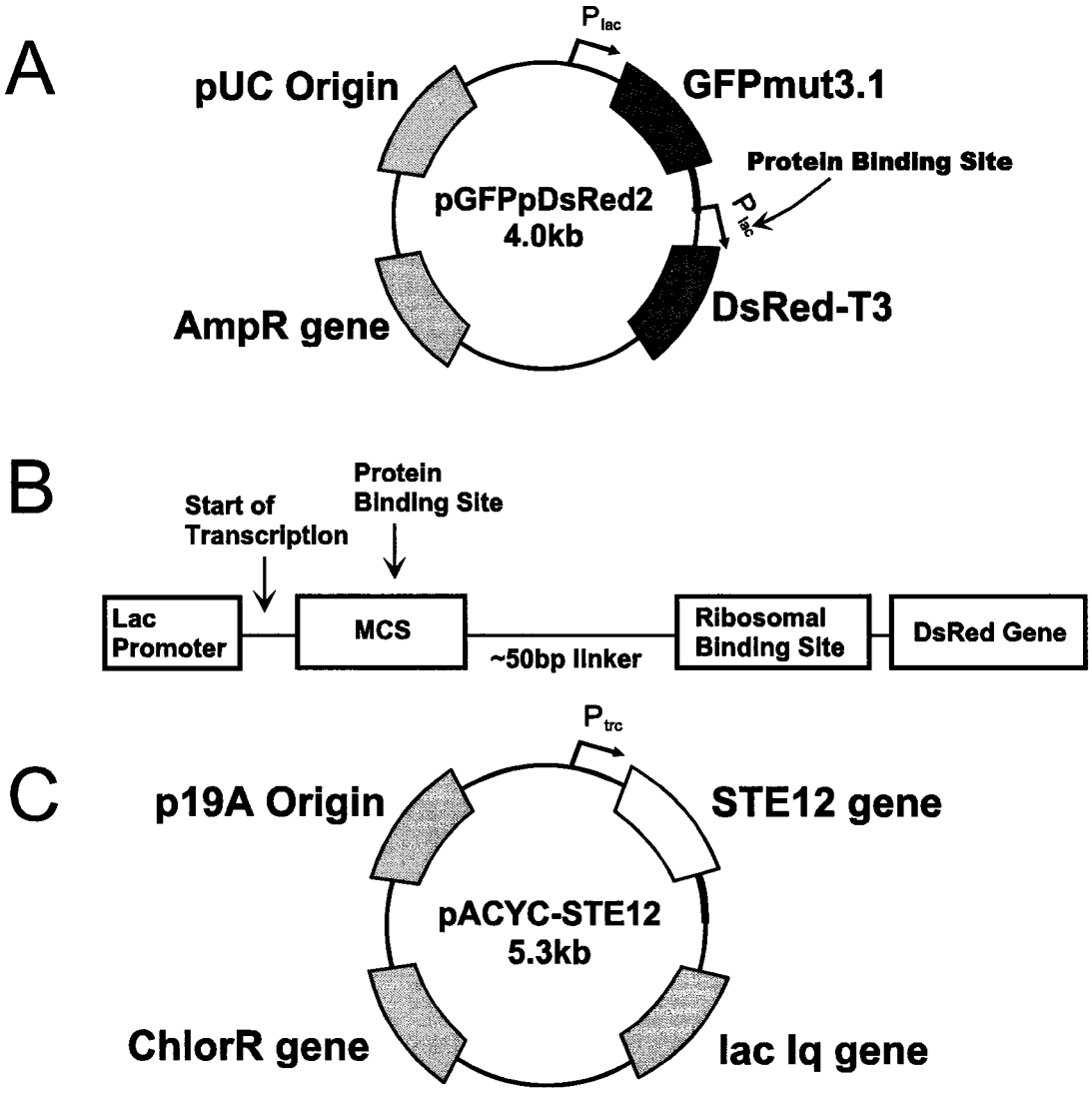



**Figure 32. Layout of pGFPpDsRed2 vector and pACYC-STE12 expression plasmid.** A) pGFPpDsRed2 is derived from pGRFP2 but contains a stop codon and a rrnB transcriptional terminator after GFP. A new promoter region has been inserted in front of the DsRed-T3 gene. B) Layout of promoter region upstream of DsRed-T3. This region contains an identical lac promoter to the promoter used for GFP. After the start of transcription point, there is a multiple cloning site (MCS) that can be used to clone in candidate protein binding sites. A 50 bp linker region separates the protein binding site from a consensus ribosomal binding site and the start codon of the DsRed-T3 gene. C) Layout of expression plasmid pACYC-STE12. The plasmid is based on pACYC184 and contains the p19A low copy origin of replication. The STE12 gene is expressed from a Trc promoter under the control of a lac Iq repressor that is also encoded on the plasmid. Protein expression can be regulated with the addition of IPTG.

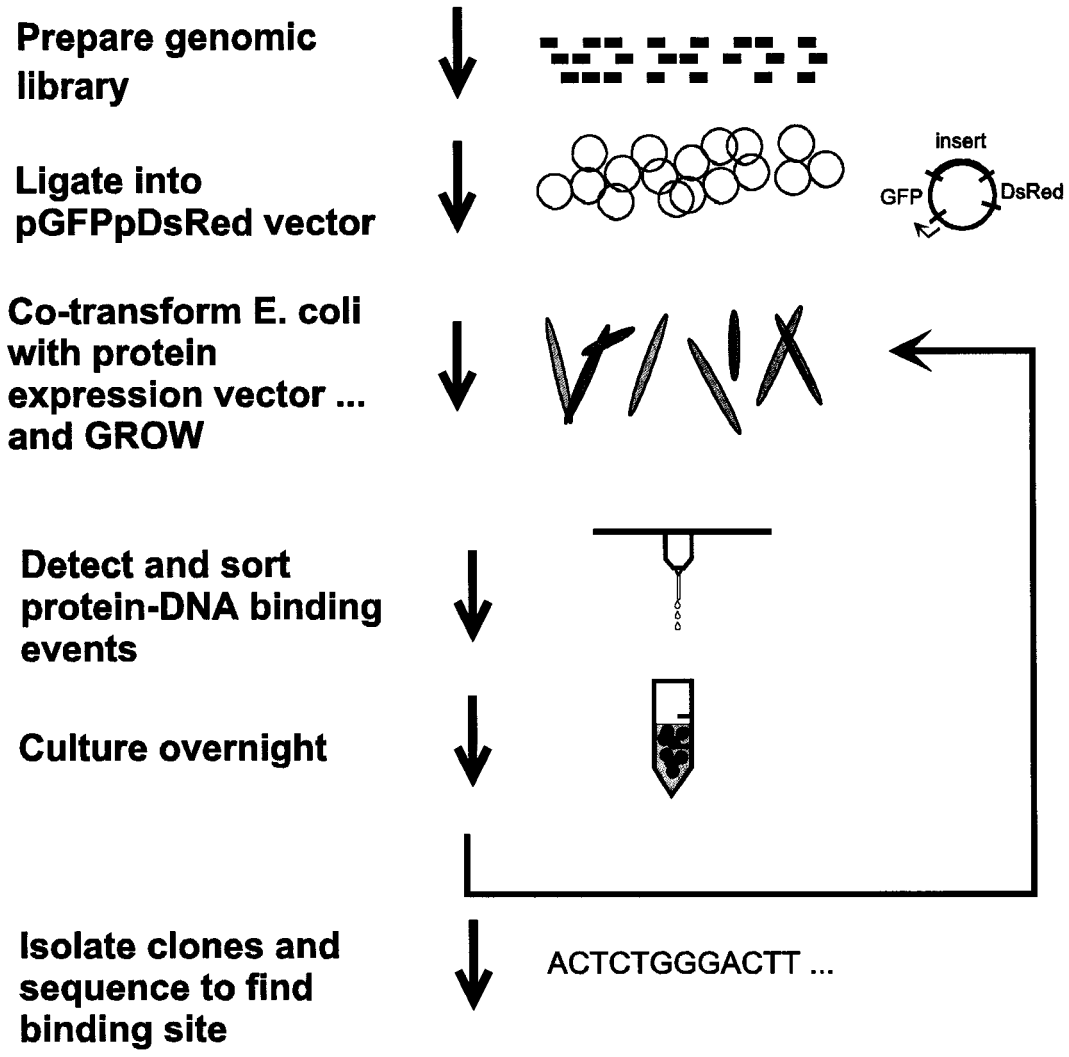



**Figure 33. Summary of proposed genomic scan for protein binding sites.** A small genomic library can be made by mechanical shearing or digesting with a frequent cutting restriction enzyme. This library can then be cloned into a pGFPpDsRed vector and transformed into *E. coli* with an expression plasmid containing the protein of interest. After culturing, cells with decreased DsRed fluorescence relative to GFP fluorescence are flow sorted and re-cultured. At this point, multiple rounds of culturing and flow sorting can be performed to further enrich for binding sites. The remaining clones can then be isolated from one another and sequenced.

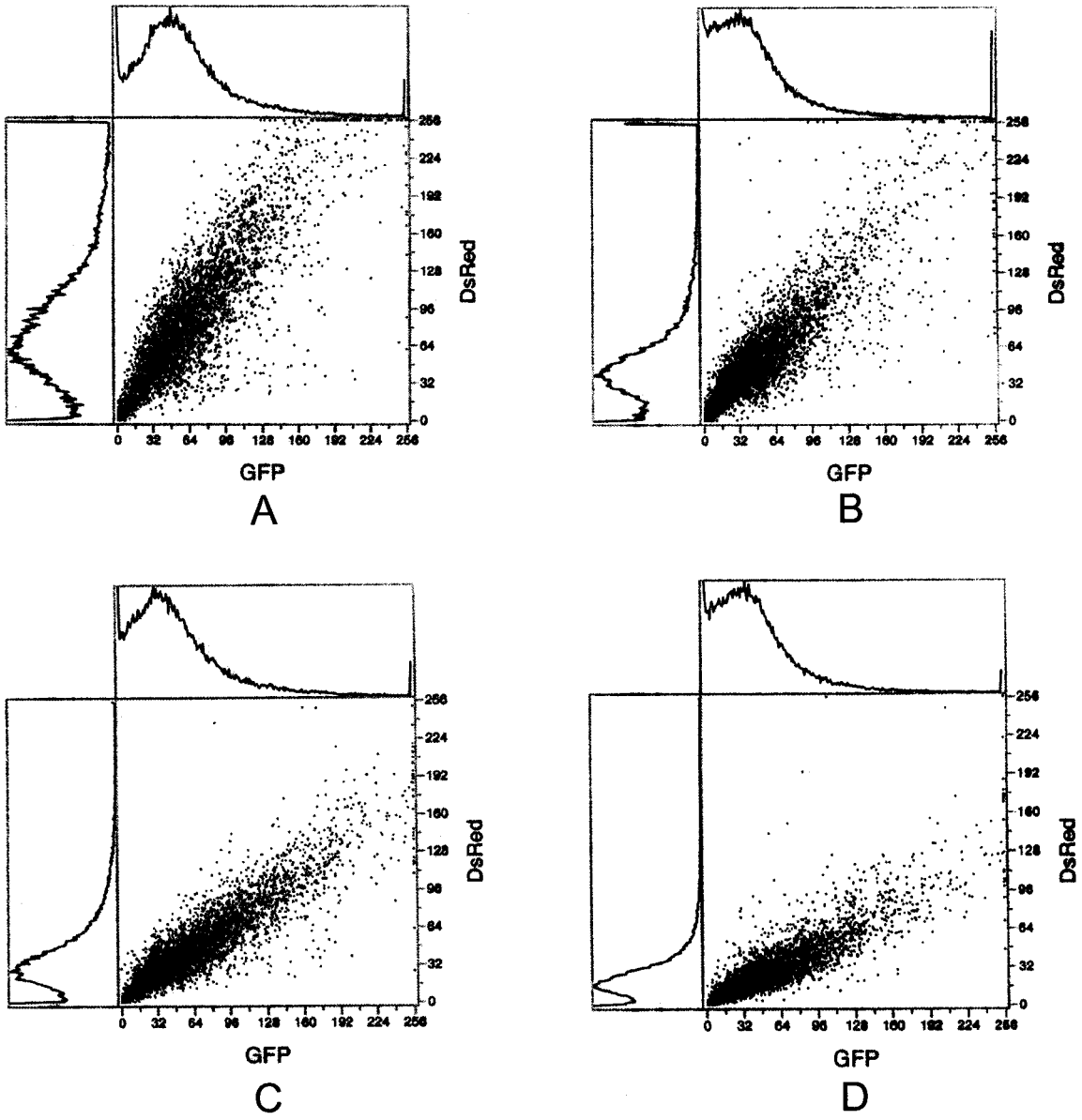



**Figure 34. Dot plots of *E. coli* cultures with pGFPpDsRed and pACYC-STE12.** All four cultures shown are expressing STE12 protein from the pACYC-STE12. The cultures differ in terms of how many STE12 binding sites are present on the pGFPpDsRed plasmid. Note the diminishment in DsRed fluorescence as the number of STE12 binding sites increases. A) This dot plot shows cells transformed with native pGFPpDsRed3 with no STE12 binding sites. B) This dot plot shows cells transformed with pGFPpDsRed3 with one cloned STE12 binding site in the "R" orientation. C) This dot plot shows cells transformed with pGFPpDsRed3 and two cloned STE12 binding site in the "FR" orientation. D) This dot plot shows cells transformed with pGFPpDsRed3 with three cloned STE12 binding site in the "FRR" orientation.

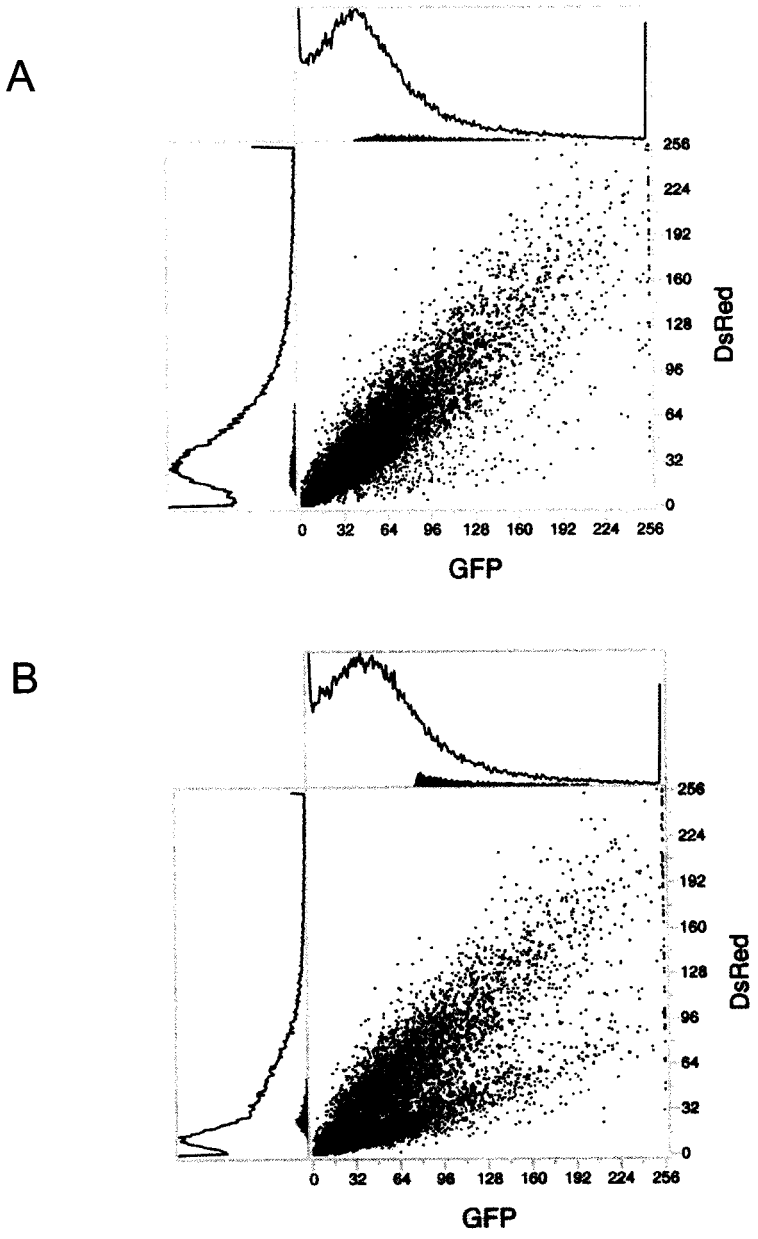



**Figure 35. Dot plots of mixed pGFPpDsRed *E. coli* cultures.** All of the cells shown are expressing STE12 protein from the pACYC-STE12. The mixed cultures differ in terms of how many STE12 binding sites are present on pGFPpDsRed3. The green dots represent cells that were sorted for this study. A) Cultures containing native pGFPpDsRed3 and pGFPpDsRed3 with one STE12 site ("R" orientation) were combined in approximately 50/50 ratio. 2.8% of total cells are marked in green as having been sorted. B) Cultures containing native pGFPpDsRed3 and pGFPpDsRed3 with three STE12 site ("FRR" orientation) were combined in approximately 50/50 ratio. 3.4% of total cells are marked in green as having been sorted.

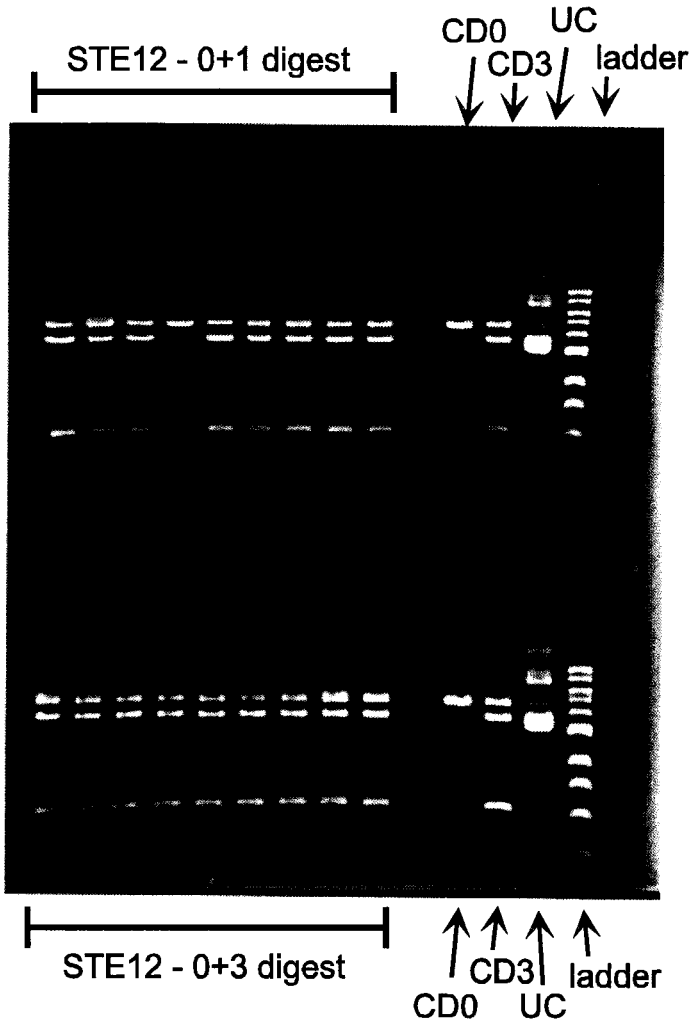



**Figure 36. Gel of NdeI digests of sorted clones from mixed pGFPpDsRed cultures.** Key: "STE12-0+1 digest" is the NdeI digest of 9 clones sorted from mixed cultures containing native pGFPpDsRed3 and pGFPpDsRed3 with one STE12 site. "STE12-0+3 digest" is the digest of 9 clones from cultures containing native pGFPpDsRed3 and pGFPpDsRed3 with three STE12 site. "CD0" is the control digest from native pGFPpDsRed3 yielding a 4.8kb product. "CD3" is the control digest from pGFPpDsRed3 with three STE 12 sites yielding 3.8kb and 1.0kb fragments (and some uncut 4.8kb). "UC" is the uncut supercoiled pGFPpDsRed3 vector. Ladder is a 1kb ladder with 500bp increments below 2kb. Note that 8 out of 9 of the sorted clones from the "STE12-0+1" pool contain STE12 binding sites. 9 out of 9 of the sorted clones from the "STE12-0+3" pool contain STE12 binding sites.

Gel image provided by Elijah Wallace and Ray Monnat.

# *Chapter 5. Fluorescence Kinetics*

## *A) Observations*

There is very little known about the kinetics of GFP fluorescence in bacterial cultures. In our own experience, when we express GFP or other fluorescent proteins from a constitutive promoter in *E. coli* under liquid culture conditions, we notice almost no fluorescence or only very dim fluorescence from individual *E. coli* cells until very late log phase to stationary phase. Cultures often grow for 12 hours or more before GFP fluorescence is observable. This is despite the fact that most newly produced GFP becomes fluorescent within 4 hours. Other groups have reported similar findings, although no further analysis or explanation has been offered in the literature for this phenomenon. For instance, in a recent paper, Zhao, et al. display data taken from a flow injection flow cytometry system that automatically samples bacteria from *E. coli* culture growing in a bioreactor.[1] One of the figures shows the cell density, forward light scatter, and GFP fluorescence over time of a growing culture of GFP-expressing *E. coli* (see figure 37). This figure shows that residual GFP fluorescence at the start of the culture decreases throughout log phase. Then, in late log phase to stationary phase, GFP fluorescence increases dramatically. However, no special note of this phenomenon was made in this paper.

To explain these results, we created a model for fluorescence kinetics in bacterial cultures. Here, we will describe computer simulations that we performed using

[1] Zhao, R., A. Natarajan, F. Srienc. (1999)."A Flow Injection Flow Cytometry System for On-Line Monitoring of Bioreactors." Biotech. & Bioeng. **62**: 609-617.

realistic parameters that we synthesized for the growth function, fluorescent protein production function, fluorescence development function, and the time constant of degradation.

### *B) Model of Fluorescence Kinetics*

We propose that the kinetics of fluorescence development in *E. coli* in culture is dependent upon a number of factors: 1) growth characteristics of the culture over time, 2) the amount of fluorescent protein each cell produces at various times during culture growth (the GFP protein production function), 3) the fluorescence development kinetics of newly produced fluorescent protein over time in the cellular environment of *E. coli*, and 4) the amount of fluorescent protein, if any, that is degraded over time.

The first factor is the growth characteristics of cultures or the change in cell density within the culture over time. *E. coli* cultures in log phase can grow at vastly differing rates depending on genetic and environmental conditions. They divide on the order of every 20-40 minutes. Fortunately, these growth rates are readily measurable by light absorbance readings or flow cytometry. We can call this growth function N(t). This is the number of cells present in culture at any given time t.

The second factor is the fluorescent protein production function. Extensive experiments have not been performed in this area, but it is expected that fluorescent protein production per cell is relatively constant under a constitutive promoter in log phase. Protein production probably decreases due to nutrient starvation as the culture

enters stationary phase. We will call this protein production function FP(t); it is the rate of fluorescent protein production per cell at any given time. Note that this protein is not fluorescent when it is first produced. Experiments could be designed to measure this FP(t) function at any point in a growing culture.

The third factor is the fluorescence development function of the fluorescent protein. This has been well studied for GFP and its mutant varieties as well as for DsRed protein. It was recently discovered that that GFP must undergo three ordered steps to become fluorescent.[2] Reid, et al. measured the kinetics of these three processes *in vitro*.[3] Protein folding occurs fairly slowly with a time constant of $k=2.44x10^{-3}$ $s^{-1}$. An intermediate step then occurs that involves cyclization and other changes of the tripeptide chromophore region of the protein. This occurs more quickly with a time constant of $k=3.8x10^{-3}$ $s^{-1}$. The final and slow step in fluorescence development is the oxidation of the cyclized chromophore. This occurs with time constant $k=1.51x10^{-4}$ $s^{-1}$. This means that half of a given amount of newly produced GFP will be fluorescent approximately 90 minutes later.

Discosoma Red Fluorescent Protein (DsRed) was recently cloned from a species of fluorescent coral.[4] It has the potential advantage of being the most red-shifted fluorescent protein isolated to date. It also shares little sequence homology with GFP so that both can be stably maintained simultaneously on plasmids in *E. coli*. However,

---

[2] Cody, C.W., D.C. Prasher, et al. (1993). "Chemical Structure of the Hexapeptide Chromophore of the Aequorea Green-Fluorescent Protein." Biochemistry **32**: 1212-1218.
[3] Reid, B.G. and G.C. Flynn (1997). "Chromophore Formation in Green Fluorescent Protein." Biochemistry **36**: 6786-6791.
[4] Matz, M.V., A.F. Fradkov, et al. (1999). " Fluorescent Proteins from Nonbioluminescent Anthozoa species." Nature Biotech. **17**:969-973.

much less is known about this protein compared with GFP. Baird et al. reported that DsRed matures through a very weak green fluorescent intermediary that appears after 7 hours and disappears to nearly zero after two days.[5] Meanwhile, the red fluorescence reaches half its maximal fluorescence after 27 hours and requires >48 hours to reach >90% of maximal fluorescence. Recent studies with mass spectra of treated DsRed fragments indicate a possible mechanism of fluorescence development.[6] It is theorized that the green fluorescent intermediate of DsRed develops by folding, cyclization, and oxidation of the chromophore, similar to what happens in GFP. Then, a loss of another water molecule and two further oxidation steps leads to the final red fluorescent form. The sheer number of steps explains why DsRed gains its red fluorescence so slowly. Because all of these kinetics parameters were measured in vitro, actual speeds in cellular environments may vary somewhat. Experiments could be designed to estimate the fluorescence development function for any fluorescent protein inside *E. coli*. We will call this function FD(t); this is the proportion of newly produced fluorescent protein that is actually fluorescent a given time t after it is produced.

The fourth factor in fluorescence kinetics in *E. coli* is the possibility of degradation of the fluorescent protein over time. Although this rate has not been measured, it is estimated to be relatively slow. In our experience, cultures stored for weeks after reaching stationary phase retain their fluorescent characteristics. In some

---

[5] Baird, G.S., D.A. Zacharias, and R.Y. Tsien. (2000). "Biochemistry, Mutagenesis and Oligomerization of DsRed, a Red Fluorescent Protein from Coral." Proc. Natl. Acad. Sci. **97**: 11984-11989.
[6] Gross, L.A., G.S. Baird, et al. (2000). "The Structure of the Chromophore within DsRed, a Red Fluorescent Protein from Coral." Proc. Natl. Acad. Sci. **97**:11990-11995.

cases, as is the case with DsRed, fluorescence actually increases over long periods of time due to increased fluorophore maturation. However, it is possible to measure this degradation factor to ensure that it is not a factor in kinetics. We can assume that this degradation occurs to all fluorescent proteins irregardless of what state they are currently in. We can assign a time constant of degradation, d, to this rate of decay.

The growth characteristics of the culture allow us to assign a number N(t) to the number of cells present in the culture at each point in time. Then, the fluorescent protein production function, P(t), gives us the rate of fluorescent protein produced by each cell on average at any given time. Therefore, the rate of total fluorescent protein produced in the culture at any given time is N(t)xP(t). In computer models, we can divide time almost as finely as we want. In a given unit of time, an amount proportional to N(t)xP(t) of new fluorescent protein will be produced. Then, in the future, this amount of fluorescent protein will become fluorescent at a rate that is determined by the fluorescence development function, FD(t). This function expresses the fraction of fluorescent protein that is actually fluorescent at a given time t after it's produced. This function will have values from 0 to 1 over time. Usually, this function is not a simple exponential decay function in which the rate of new fluorescence produced is directly proportional to the amount of remaining non-fluorescent protein. Because fluorescent proteins enter many states during their courses of fluorescence development, the fluorescence development function is often non-linear. Therefore, our model must keep track of each packet of k x N(t) x P(t) fluorescent protein throughout the lifetime of the culture to determine how much of it

becomes properly processed to its full fluorescent form. At a given time t, the total amount of fluorescence in the culture TF(t) is given by:

$$TF(t) = \sum_{n=1}^{n=t-1} k \bullet N(n) \bullet P(n) \bullet FD(t-n) .$$

If there is a significant protein turnover rate, then we can approximate it roughly as an exponential decay function. In other words, in each segment of time, an amount of fluorescent protein is lost that is proportional to the total amount of fluorescent protein. We will assume that the rate of degradation of fluorescent protein is constant regardless of what state it is in. If we assume an exponential decay time constant d, then we get:

$$TF(t) = \sum_{n=1}^{n=n-1} k \bullet N(n) \bullet P(n) \bullet e^{-d(t-n)} FD(t-n) .$$

However, this time constant d may be so large that it becomes irrelevant to cultures growing during the period of one or two days. In that case, we can revert back to the first TF(t) equation. In any case, we can then define the average fluorescence per cell F(t) in the culture.

$$F(t) = \frac{\sum_{n=1}^{n=t-1} k \bullet N(n) \bullet P(n) \bullet e^{-d(t-n)} FD(t-n)}{N(t)}$$

This quantity could be measured directly in a flow cytometer for a given cell type.

C) ***Results of Simulation***

Each time index for t will represent one minute of time. We specified a plausible function N(t) for the growth function (see figure 38A). At t=0 at the beginning of log phase the culture has 10,000 cells. Doublings continue every 30 minutes until t=540 (9 hours), when growth begins to level off. By t=720 (12 hours), the culture is in stationary phase and growth has stopped.

The fluorescent protein production function P(t) was estimated as having a constant value of 1 until about t=540 (9 hours), when growth of the culture also began to level off (see figure 38B). This assumes that cells are producing fluorescent protein at the same rate per cell throughout log phase. After t=540, the fluorescent protein production capability is assumed to decrease as a result of nutrient depletion. This is probably an accurate assessment, although this area requires experimental testing.

The fluorescence development function FD(t) was estimated based on published time constant measurements for GFP.[7] This function is shown in figure 39. We modeled GFP as it progressed through four states: raw, folded, cyclized, and

[7] Reid, B.G. and G.C. Flynn (1997). "Chromophore Formation in Green Fluorescent Protein." Biochemistry **36**: 6786-6791.

fluorescent using the Reid, et al. kinetics parameters. 50% of the fluorescent protein is fluorescent after 90 minutes. 90% of the fluorescent protein is fluorescent after 260 minutes.

We assumed for this simulation that no significant degradation of fluorescent protein occurs. Therefore, the degradation time constant d is equal to ∞, and the term $e^{-d(t-n)}$ goes to 1. It may be possible to eventually determine this constant, but it is expected to be large amount of time.

Given these four parameters, we simulated the fluorescence kinetics for this culture from t=0 to 840 (14 hours) (see figure 40). The fluorescence per cell increases in the first 200 minutes, because we are assuming that the starting cells are non-fluorescent at t=0. As GFP is produced initially, and it folds, some fluorescence develops. However, large amounts of fluorescence can not accumulate due to the high rate of division that continuously dilutes the amount of fluorescent GFP. The fluorescence reaches a steady state level that is maintained throughout log phase. As the rate of division begins to slow starting at t=540 (9 hours), we see a rapid increase in GFP fluorescence. In the long run, we predict that GFP production drops to zero, and cells stop dividing. This stabilizes the level of GFP fluorescence to another steady state value. This pattern of low GFP fluorescence during log phase followed by a sudden increase in late log phase to stationary phase largely fits what we observe by fluorescence microscopy in actual cultures.

We further focused on the fluorescence kinetics of cultures in log phase. We noticed that during log phase, the fluorescence per cell reaches a steady state value

that does not change until the replication rate changes. We calculated a large number of these steady state values with respect to the doubling time (see figure 41). We kept the fluorescence development function [FD(t)] the same as in figure 40. Also, we set the protein production function [P(t)] to a constant value during this log phase simulation. This resulted in the discovery of an approximately linear relationship between doubling time and steady state fluorescence values. The faster the replication rate, the lower the steady state fluorescence value is. This makes sense, because more dilution is occurring relative to the amount of protein production.

We also simulated the steady state fluorescence values for cells relative to the speed of protein folding (see figure 42). We kept the bacterial doubling time constant at 30 minutes. We also kept the protein production function constant for this simulation. We only altered the rate of fluorescence development from each hypothetical florescent protein variant. We modeled the fluorescence kinetics as a simple exponential decay from the non-fluorescent form to the fully fluorescent form. The time constant of fluorescent protein processing corresponds to the amount of time a protein requires to reach 63.2% of its full fluorescence. The results of this simulation show that there is an inverse relationship between the fluorescent protein processing time constant and the steady state fluorescence per cell. In other words, as fluorescent proteins are expressed that have slower rates of fluorescence development, the overall steady state fluorescence is lower during log phase. This makes sense, since slower developing fluorescent proteins must wait longer before

they exhibit fluorescence. This allows more dilution to occur by bacterial replication during this time period.

### *D) Conclusion*

Putting reasonable parameters into this model, it becomes clear that the levels of fluorescence in bacteria are being artificially lowered by their high rate of replication. A given amount of fluorescent protein is produced during culturing; however, by the time that fluorescent protein has folded and undergone modifications to become fluorescent, it has been diluted by several bacterial divisions. The result is that fluorescence per cell stays relatively low during periods of rapid growth. The amount of fluorescent protein produced can never catch up with the rate of cell division. When replication finally slows at the end of log phase, fluorescence levels increase dramatically, because the same amount of fluorescent protein is being produced with less dilution.

Understanding these processes has implications for a wide variety of fields. There are many groups using GFP and other fluorescent proteins in bacterial and yeast systems. In any system in which the cell cycle time is fast compared to the fluorescent protein processing time and the fluorescent protein production rate, a dilution effect can occur that reduces intracellular fluorescence levels. If fluorescence from fluorescent proteins is being quantitated to reveal biological information, then the division time of the cells must also be factored in. Otherwise, gross errors could result. This phenomenon is also relevant beyond the scope of fluorescent proteins.

Presumably, any protein growing in bacteria or other cells would fall prey to these effects. The protein content per cell depends heavily on the rate of replication of these cells. This has implications for experimenters who rely on cell counts to estimate total protein yields. Industrial applications involving protein manufacturing would also be greatly effected. Simply taking an optical density reading might provide a good measure of the number of cells in a culture, but would not necessarily provide a good measure of total protein content.

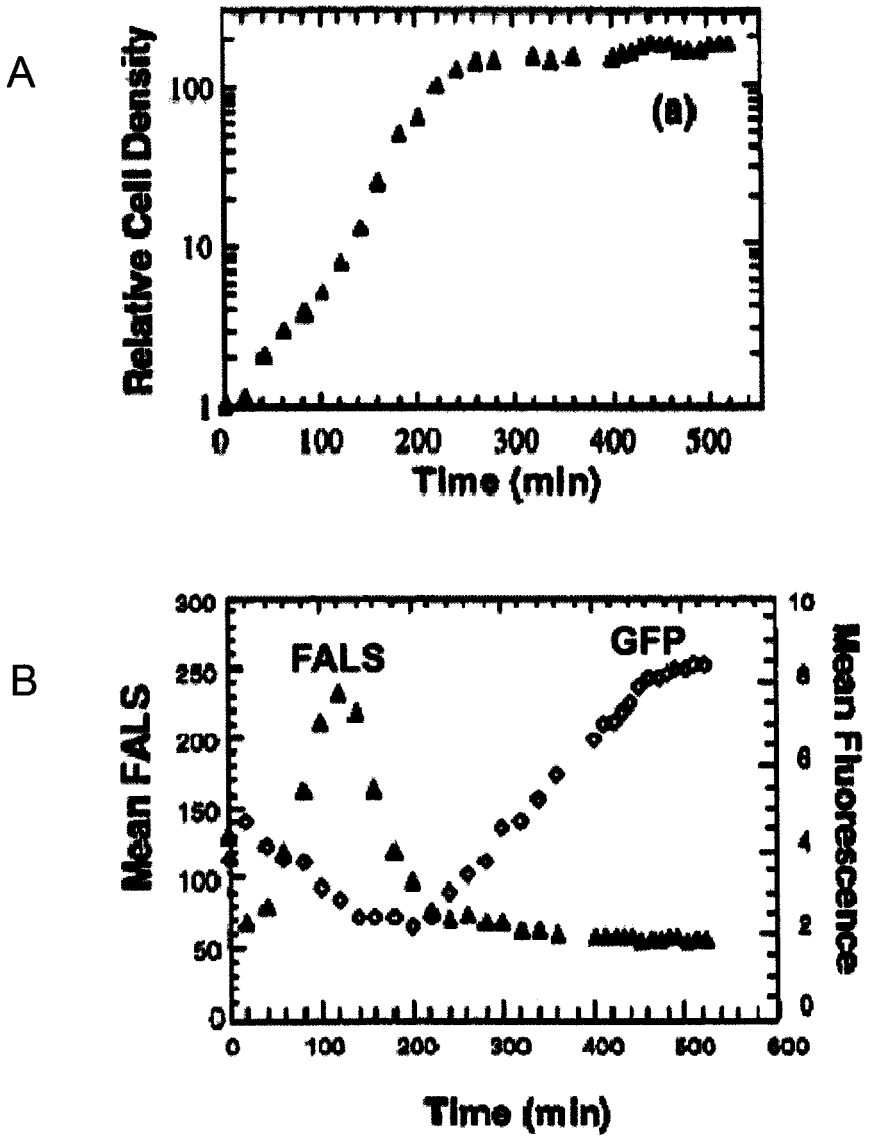
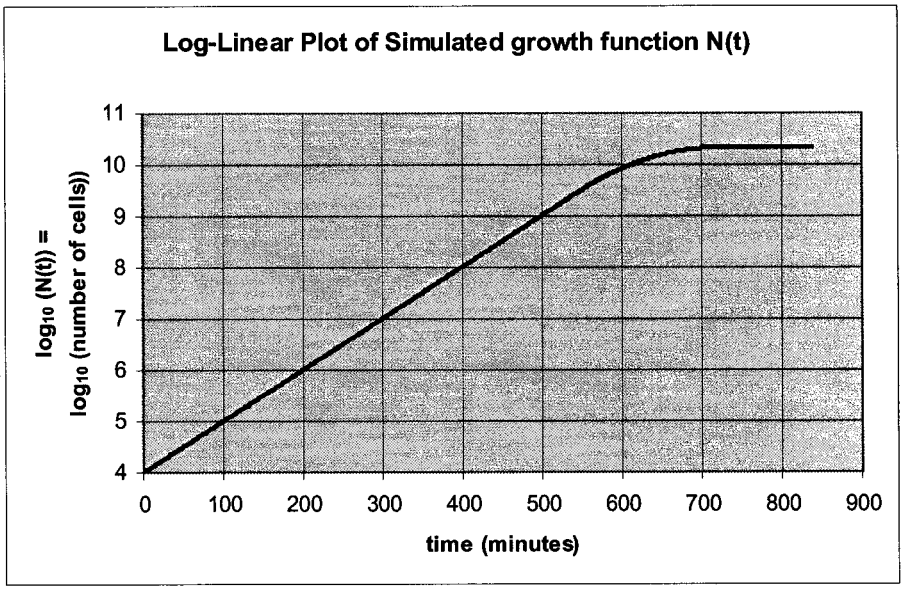



**Figure 37. Flow cytometric analyses of growing GFP-expressing *E. coli* in culture.** A)The cell density of the culture was estimated by running cells undiluted through a flow cytometer. Notice the start of observation in log phase at t=0, with the start of stationary phase around t=250-300. B) In the same culture, forward light scatter (triangles) and GFP fluorescence (diamonds) are quantitated by flow cytometry. Because the initial cells already contained GFP, some residual GFP fluorescence can be seen at t=0. This fluorescence decreases throughout log phase and only begins to increase again in late log phase to stationary phase.

Adapted from Zhao, et al. 1999.

A




B

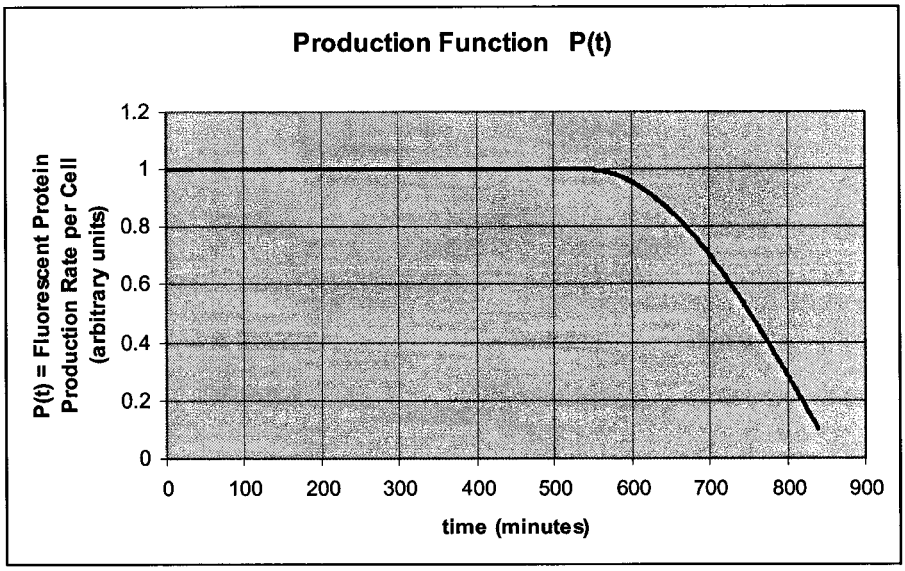



**Figure 38. Plots of estimated bacterial growth and fluorescent protein production.** A) A plausible growth function N(t) for a bacterial culture is shown. Exponential growth continues for 9 hours and then begins to level off. It is completely in stationary phase at t=720 minutes (12 hours). B) A simulated production function P(t) is shown. This function represents the rate of raw fluorescent protein produced by each cell at any given time during the lifetime of the culture.

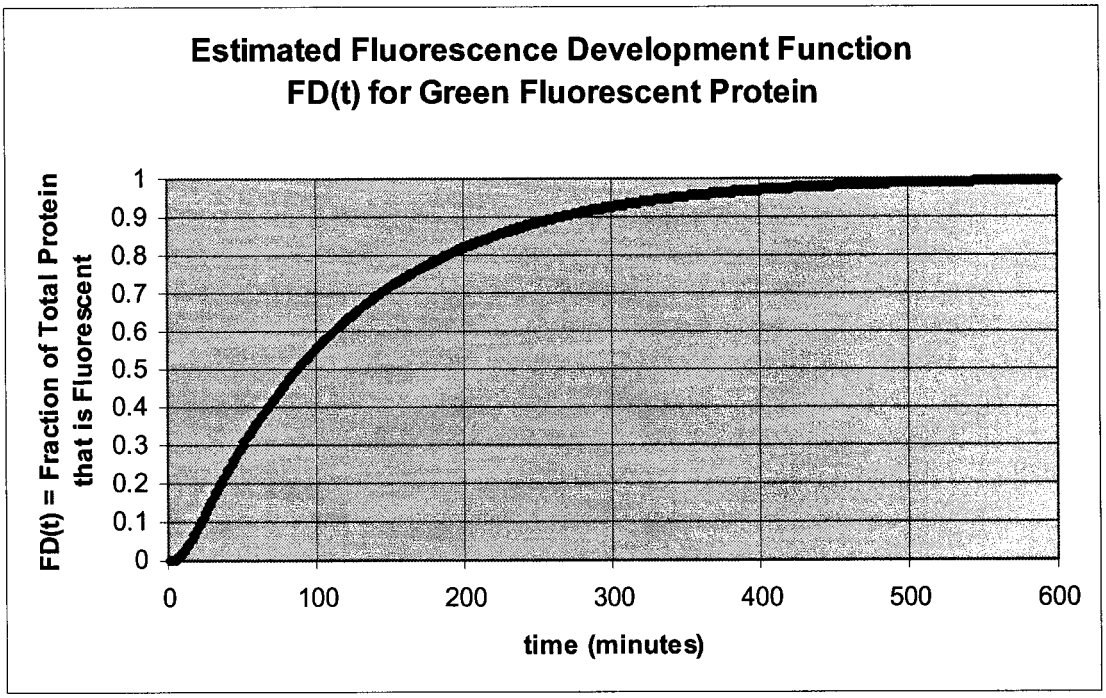



**Figure 39. Plot of estimated fluorescence development function.** These values are based on time constants measured by Reid, et al. We simulated an amount of raw fluorescent protein as it progresses through the following states: raw, folded, cyclized, and fully fluorescent. Half of the fluorescent protein is fluorescent after 90 minutes. 90% of the fluorescent protein is fluorescent after 260 minutes.

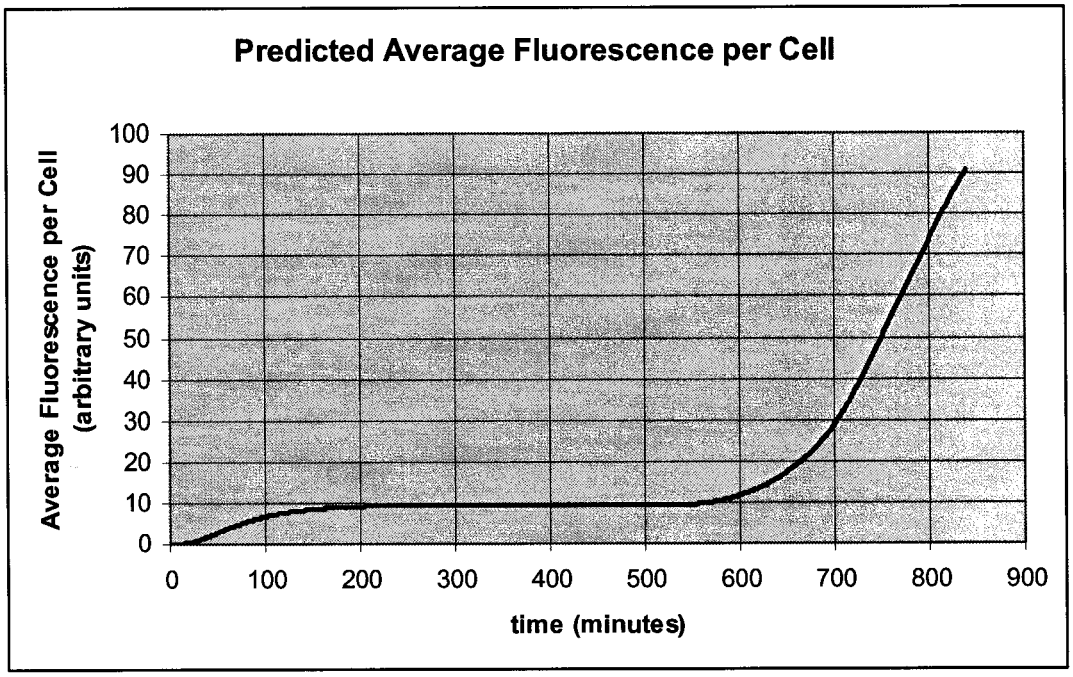



**Figure 40. Plot of predicted average fluorescence per cell.** Given the functions illustrated in figures 23-24, the predicted average fluorescence per cell is shown above. These values were predicted by our model described in the "Background" section on fluorescence kinetics.

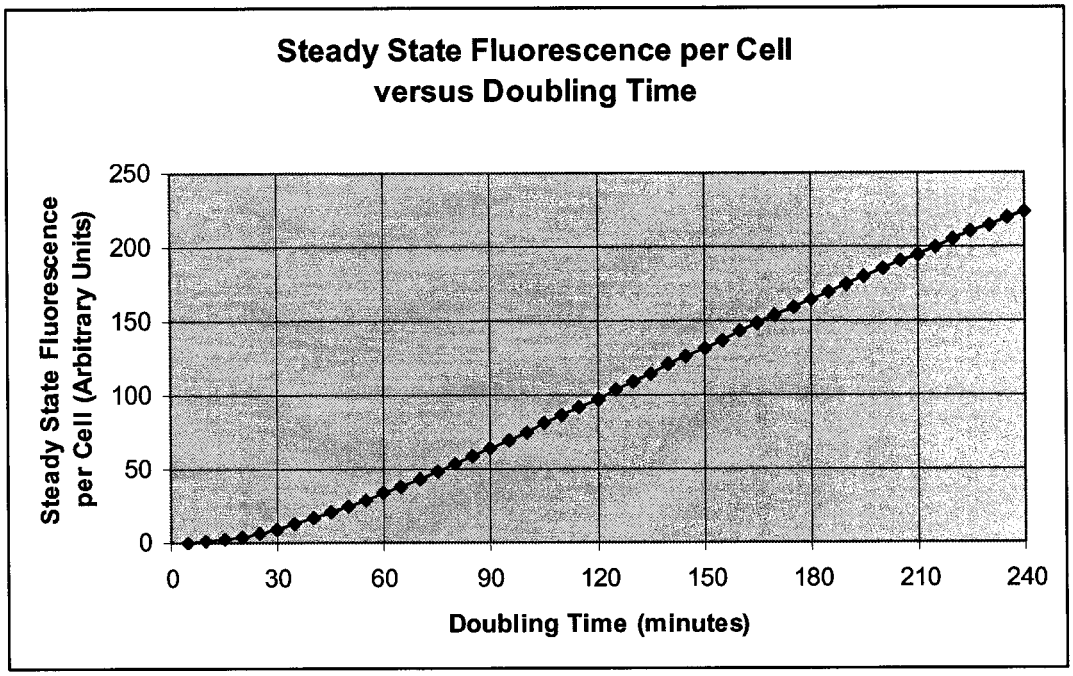



**Figure 41. Plot of steady state fluorescence per cell versus doubling time.** This simulation was performed with numerous virtual cultures, each with a different rate of replication. Doubling times from five minutes to 240 minutes were simulated at 5 minute intervals. The average fluorescence per cell was plotted after it reached steady state while still in log phase. The protein production function was presumed to be constant during the course of each simulation. The fluorescence development function was the same as the theoretical GFP function shown in figure 28.

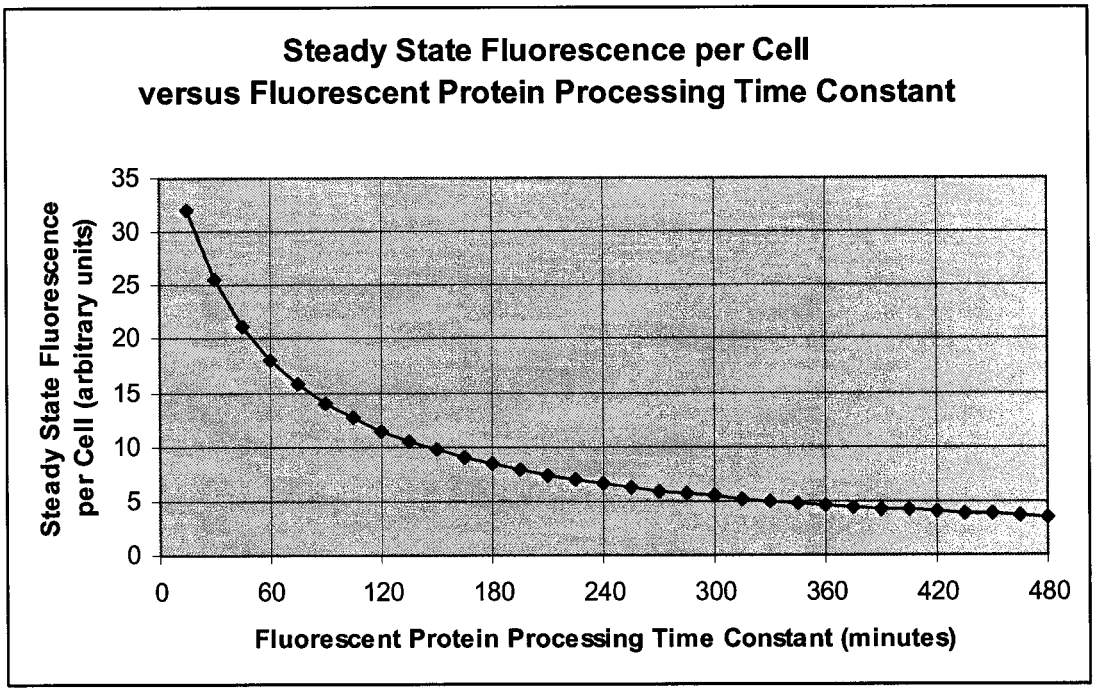



**Figure 42. Plot of steady state fluorescence versus fluorescent protein time constant.** This simulation was performed with numerous virtual cultures, each with a different time constant for fluorescent protein processing. Each fluorescent protein was approximately modeled as having an exponential decay towards full fluorescence. The time constant is the amount of time needed for a given amount of fluorescent protein to reach ~63.2% of its full fluorescence. Time constants were simulated from 15 minutes to 480 minutes at 15 minute intervals. The average fluorescence per cell was plotted after it reached steady state while still in log phase. The doubling time was set at 30 minutes. The protein production function was presumed to be constant during the course of each simulation. The GFP fluorescence development function shown in figure 28 roughly correlates with an exponential decay function with time constant of 130 minutes.

# *Chapter 6. RecA Independent Recombination*

## *A) Observations*

During our experiments with pBGFP, the cloning vector with a blue fluorescent protein gene translationally fused to a green fluorescent protein gene, we observed a high frequency of rearrangements. In this chapter, we will show the evidence for recombination in this plasmid, and explain how this may be occurring in a RecA negative strain. Recombination is a potentially critical problem that can occur with certain vector designs.

A 2-3 kb size-selected Adenovirus-2 genomic library was cloned into pBGFP. This library was transformed into DH10B cells, a RecA minus strain. A culture was started and grown for 24 hours at 37 degrees C. Single BFP+/GFP- cells were flow sorted into 96-well plates containing LB medium. After growing overnight, cultures with growth were picked into PCR reactions to amplify the inserts. Out of twenty PCR reactions, 13 wells contained a strong signal for an amplified insert in the expected 2-3kb range (see figure 14). The remaining seven reactions resulted in either no PCR amplification or very weak amplification.

To further examine these seven clones, they were grown in larger volumes, and plasmid DNA was isolated with a commercial miniprep kit. Restriction analysis was performed with an Afl II/Not I digest, a Afl II / Hae II digest, and a single Afl II digest (see restriction sites on figure 11). In the native pBGFP plasmid, we expect

fragments of length 1059bp and 2750bp for the Afl II/Not I digest. Clones 6 and 7 do not contain a version of pBGFP without a NotI site and expect only one 3.8kb fragment. If an insert is present, it would contribute to the 1059bp fragment (or 3.8kb fragment for clones 6 and 7). If additional restriction sites are present in the insert, then the 1059bp+insert (or 3.8kb+insert) fragment would be further subdivided. Similarly, the Afl II/Hae II digest should yield 1629bp and 2180bp fragments in the native vector. An insert would add to the 2180bp fragment. The single Afl II digest cuts only once and should yield a 3.8kb fragment for the native vector. We found that Afl II does not cleave in any of the inserts and the digest products provide a very accurate estimate of total plasmid size. We estimated the fragment sizes of the digest products on agarose gels. Note that some overlapping fragments may only appear as a single band on the gels. We also observed the cultures under fluorescence microscopy to evaluate their fluorescence profiles. These results are summarized below.

**Table 7: Restriction analysis and fluorescence of 7 clones that did not PCR amplify**

| Clone | Afl II/Not I Digest Products | Afl II / Hae II Digest Products | Afl II Digest (total length) | Fluorescence Profile |
|---|---|---|---|---|
| 1 | 2.8 kb | 1.6kb, 1.2kb | 2.8 kb | No fluorescence |
| 2 | 2.9 kb | 1.6kb, 1.4kb | 2.9 kb | BFP+ / GFP - |
| 3 | 2.9 kb | 1.6kb, 1.4kb | 2.9 kb | BFP+ / GFP - |
| 4 | 2.7 kb, 2.3 kb, 700 bp, 300bp | 1.9kb, 1.6kb, 1.3kb, 900bp, 400bp | 5.8 kb | BFP+ / GFP - |
| 5 | 2.9 kb | 1.6kb, 1.3kb | 2.9 kb | No fluorescence |
| 6 | 3.8 kb, 1.3 kb, 300 bp | 1.6kb, 400bp | 6 kb | No fluorescence |
| 7 | 3.8 kb, 1.2 kb, 900 bp, 300 bp | 1.6kb, 1.2kb, 900bp, 500bp,300bp, 200bp | 7 kb | BFP+ / GFP - |

Four of the seven clones (#1,2,3, and 5) have an estimated total length of approximately 2.8-2.9 kb. This is considerably smaller than the 3.8kb starting size of the pBGFP vector. This strongly points to a deletion process that is occurring within this vector. If the BFP and the GFP gene are thought of as direct repeats, it is conceivable that one of these fluorescent proteins could have been effectively deleted. The length of each fluorescent protein gene is 717bp. This is tantalizingly close to the estimated 900bp that is missing. The three clones (#5,6, and 7) with large total

plasmid sizes are most likely pBGFP vector with successfully integrated inserts. We can infer this from the additional size of fragments containing the EcoRV cloning site. PCR products may not have been obtained from these three clones due to inefficiencies in amplification due to the content of the inserts. Clone 6 probably lacks fluorescence due to a small deletion caused by exonuclease activity starting at the EcoRV cloning site. An insert could then have ligated into this shortened version of pBGFP.

We sequenced the four shortened clones (#1,2,3, and 5) in the forward and reverse directions using primers that flank the entire BFP-GFP construct. In clones 1 and 5, we detected a 945bp and a 903bp deletion respectively encompassing the lac promoter and BFP gene all the way to the EcoRV cloning site. This is most likely due to exonuclease activity that digested the plasmid from the EcoRV site during cloning. In clones 2 and 3, the GFP gene has been deleted, and the BFP gene has been joined cleanly to the DNA directly after the original GFP gene. The stop codon from GFP has been added directly to the end of the BFP gene. Because the degree of sequence similarity between BFP and GFP is so high, it is difficult to tell whether GFP was simply deleted or a hybrid molecule formed between BFP and GFP. The last sequence difference between the two genes occurs at codon 145. It appears that the single remaining fluorescent protein in clones 2 and 3 is identical to BFP including at codon 145. A diagram of the rearrangement from native pBGFP to recombinant clones 2 and 3 is shown in figure 43.

Despite the fact that these clones were grown in a RecA minus host, there is clear evidence that a rearrangement involving the BFP and GFP genes has occurred. By looking at theories for RecA independent recombination in the literature, we can infer what mechanisms may have been involved in causing the deletions in clones 2 and 3.

### *B) Background*

Recombination is currently a very active area of research. One reason is recombination's inevitable involvement with the understanding of replication. Double strand breaks can occur due to ionizing radiation or as a natural consequence of DNA replication on a chemically flawed template. Unless this double stranded break can be repaired, the lesion is inevitably lethal as DNA replication is normally halted at that site. Recombination is the necessary tool that allows repair of the damaged DNA region as well as helping to reform the replication fork.[1] This function of recombination is remarkably conserved, and RecA-like proteins are thought to exist in almost all prokaryotes and eukaryotes.[2]

Recombination also serves the important evolutionary function of allowing the interchange of genetic materials. In prokaryotes, there is often extensive recombination between genes and homologues after conjugation. In eukaryotes, the existence of homologous chromosomes allows for recombination to create

---

[1] Asai, T., D.B. Bates, and T. Kogoma. (1995). "DNA Replication Triggered by Double-Stranded Breaks in *E. coli*: Dependence on Homologous Recombination Functions." Cell **78**: 1051-1061.
[2] Kowalczykowski, S.C. and A.K. Eggleston. (1994). "Homologous Pairing and DNA Strand-Exchange Proteins." Annu. Rev. Biochem. **63**: 991-1043.

chromosomes with new combinations of constituent gene alleles. Recombination is necessary to provide the genetic diversity necessary for selection. Otherwise, a small group of chromosomes would be propagated with very few changes from generation to generation.

Recombination is relevant to this discussion due to its implications in altering tandemly repeated sequences on plasmid DNA. In these situations, recombination between direct repeats can generally be grouped into two primary divisions: RecA dependent and RecA-independent recombination.

RecA recombination is a fairly well studied system involving over 25 known contributory proteins. However, almost all pathways in the RecA system require the RecA protein. RecA recombination is thought to have little dependence on the distance between recombining elements. In fact, RecA recombination is thought to be one of the few mechanisms within *E. coli* for both interstrand and intrastrand recombination. In conjugation studies, RecA mutants had greater than five orders of magnitude less homologous genetic recombination.[3] The RecA protein is an extremely complex protein that possesses ATPase, coprotease, DNA renaturation, and DNA strand exchange activities.[4] RecA is also thought to be the initiating protein in the SOS pathway that upregulates the production of many DNA repair enzymes in response to DNA lesions.[5] RecA recombination acts at sites of double strand breaks

---

[3] Low, B. (1968). "Formation of merodiploids in matings with a class of Rec- recipient strains of Escherichia coli K12." Proc. Natl. Acad. Sci. **60**:160-167.
[4] Kowalezykowski, S.C. (2000) "Initiation of Genetic Recombination and Recombination-Dependent Replication." Trends in Biochem. Sci. **25**: 156-165.
[5] McEntee K. (1977). "Protein X is the product of the recA gene of Escherichia coli." Proc. Natl. Acad. Sci. **74**: 5275-5279.

that can form spuriously but are more commonly the result of aborted replication at sites of damaged bases. A brief explanation of the main RecA-dependent recombination and repair process follows.[6] Note that many accessory protein interactions have been left out for brevity.

At the site of the double strand break, the RecBCD complex binds. The double stranded DNA is unwound and the exonuclease activity of the complex chews both strands back, usually until a $\chi$ site is reached. These $\chi$ sites are specialized hotspots for recombination that are scattered throughout the *E. coli* genome. The $\chi$ site is recognized by the RecBCD complex, and the nuclease activity is moderated until a 3' overhang is achieved (see figure 44B). Multiple RecA proteins then attach to this 3' overhang and cause it to infiltrate into double stranded DNA that is homologous to that 3' region, displacing one strand (see figure 44C). This is the D-loop structure. Then, replication and strand displacement leads to the formation of the Double Holliday junction (see figure 44D). The branch migration complex RuvAB allows movement of the two crossover points to span the region of DNA to be exchanged between the two homologous regions. Finally, a resolvase, RuvC, is involved in cutting the two crossover points to create two possible sets of recombined products. Ligase patches any remaining nicks. (see figure 44E).

In comparison, RecA-independent recombination between tandem repeats is poorly understood. Traditionally, it was believed that RecA formed the core of almost

---

[6] Kowalezykowski, S.C. (2000) "Initiation of Genetic Recombination and Recombination-Dependent Replication." Trends in Biochem. Sci. **25**: 156-165.

all recombination in *E. coli* and that any RecA-independent recombination was several orders of magnitude less efficient. However, RecA-independent recombination has been increasingly recognized as a major contributor to rearrangements. In one experiment, approximately 1/3 of the deletions in a 787 bp direct repeat were attributed to RecA-independent sources.[7] Mutation analysis indicates that the RecA-independent recombination process is independent of most of the enzymes involved in the RecA dependent pathway including RecA, RecBCD, RuvAB, and RuvC.[8] There is evidence of some dependence on Exonuclease I, type I topoisomerase, and topoisomerase III.[9,10,11]

Several hallmarks of RecA-independent recombination exist that distinguish it from RecA-dependent recombination. 1) A proximity effect exists in RecA-independent recombination so that when the repeated sequences are separated by more than 100bp of intervening sequence, the frequency of this form of recombination decreases rapidly.[12] Contrary to this, RecA-dependent recombination has no proximity effect and can even recombine sequences present on two separate DNA molecules. 2) RecA-independent recombination has less dependence on homology, and it has been shown that repeats shorter than 15bp can recombine

---

[7] Lovett, S.T., P.T. Drapkin, et al. (1993). "A Sister-Strand Exchange Mechansim for recA-Independent Deletion of Repeated DNA Sequences in *Escherichia coli*." Genetics **135**: 631-642.
[8] Ibid.
[9] Allgood, N.D. and T.J. Silhavy. (1991). "*Escherichia coli* xonA (sbcB) Mutants Enhance Illegitimate Recombination." Genetics **127**: 671-680.
[10] Whoriskey, S.K., M.A. Schofield, and J.H. Miller. (1991). "Isolation and characterization of *Escherichia coli* mutants with altered rates of deletion formation." Genetics **127**:21-30.
[11] Yi, T.-M., D. Stearns, and B. Demple. (1988). "Illegitimate recombination in an *Escherichia coli* plasmid: modulation by DNA damage and a new bacterial gene." J. Bacteriol. **170**:2898-2903.
[12] Bi, X. and L.F. Liu. (1996). "A Replicational Model for DNA Recombination between Direct Repeats." J. Mol. Biol. **256**: 849-858.

effectively.[13] However, there was very little RecA-dependent recombination detectable with repeats 150bp or shorter.[14,15]

The mechanisms of RecA-independent recombination have not been completely solved, but there are several competing theories explaining how this mode of recombination works. There is even evidence that multiple methods of recombination work together to contribute to the observed occurrences of recombination.[16] The most common theory for RecA-independent recombination is the replication misalignment model.[17] In this model, a deletion or an expansion can occur (see figure 45). In the case of a deletion, the template strand bunches together during replication and the nascent strand contains only one copy of the repeated region. It is important to note that this single copy is actually a hybrid made up of partial copies of two templates. In the case of expansion, the nascent strand bunches together after the two repeats are replicated. Then, the second repeat is copied again, resulting in three copies of the repeated region (triplet repeat). None of these three copies are hybrids. It is thought that a specific helicase could aid in melting apart the two strands in both cases. The replication misalignment model is attractive, because it is consistent with the need for

---

[13] Bi, X. and L.F. Liu. (1996). "A Replicational Model for DNA Recombination between Direct Repeats." J. Mol. Biol. **256**: 849-858.
[14] Bi, X. and L.F. Liu. (1994). "RecA-Independent and RecA-dependent Intramolecular Plasmid Recombination" J. Mol. Biol. **256**: 849-858.
[15] Lovett, S.T., P.T. Drapkin, et al. (1993). "A Sister-Strand Exchange Mechansim for recA-Independent Deletion of Repeated DNA Sequences in *Escherichia coli*." Genetics **135**: 631-642.
[16] Bzymek, B. and S. Lovett. (2001). "Instability of Repetitive DNA Sequences: The Role of Replication in Multiple Mechanisms." Proc. Natl. Acad. Sci. **98**: 8319-8325.
[17] Albertini, A.M., M. Hofer, M.P. Calos, J.H. Miller. (1982). "On the Formation of Spontaneous Deletions: the Importance of Short Sequence Homologies in the Generation of Large Deletions." Cell **29**:319-328.

only small regions of homology as well as the dependence on proximity between the two repeats.

There are several other models as well that will be described very briefly. In the sister-chromosome exchange (SCE) model, it is thought that unequal crossovers between sister strands of a replicating plasmid could create dimerized plasmids containing both a deletion and triplet repeat simultaneously.[18] This is shown in figure 46. If strand realignment accompanies sister strand crossovers, then a dimer consisting of a deletion and the original duplication will result. Both of these dimerized plasmid species were found in experiments by Lovett, et al. Figure 47 shows the unequal crossover event occurring at a replicational fork along with the nicking and synthesis that resolves it. According to this model, the deletion product and some of the elements in the triplet repeat or duplication could become hybrids, although not necessarily so. Since the Lovett, et al. experiments involved perfect repeats, it was not known if hybrids were created or not.

A replicational model proposed by Bi, et al. allows for the production of deletions and dimerized plasmids containing triplication/deletion and duplication/deletion dimers that Bi, et al. observe.[19] These products are similar to what can be produced using the replicational slippage or the sister chromosome exchange models. Bi, et al.'s model relies on replicational slippage coupled with endonuclease and ligase activity to connect normally unrelated strands and to delete

---

[18] Lovett, S.T., P.T. Drapkin, et al. (1993). "A Sister-Strand Exchange Mechansim for recA-Independent Deletion of Repeated DNA Sequences in *Escherichia coli*." Genetics **135**: 631-642.
[19] Bi, X. and L.F. Liu. (1996). "A Replicational Model for DNA Recombination between Direct Repeats." J. Mol. Biol. **256**: 849-858.

out single stranded loops. However, due to the number of unlikely events that need to occur, this model is somewhat discounted. The Bi, et al. model also departs from the replicational slippage and sister chromosome exchange model by suggesting that none of the repeated sequences become hybrids during recombination. In other words, in the Bi, et al. model, there is no recombination occurring in the middle of repeat sequences that could lead to hybrids comprised of combinations of the two repeated sequences.

### *C) Conclusions*

Contrary to common wisdom, RecA independent recombination can occur at high frequencies in some special circumstances. Certainly, constructing a vector with large areas of sequence similarity such as a blue fluorescing mutant gene of GFP (BFP) and a standard GFP gene provides a large target for recombination. Each gene represents 714bp of similar sequence with only 11bp difference between the two. The second factor aiding recombination may be the small distance of intervening space between the two genes. Only 89bp of non-repeated sequence exists between the BFP and GFP genes.

In our experience, two clones (clones 2 and 3) out of 20 sorted clones contained the type of deletion rearrangement predicted by models of RecA independent recombination. Despite the small sample size, 10% of all clones represents a significant fraction of rearranged clones. Experimental conditions may also have led to this high frequency of recombination. The long 24 hour incubations at 37 degrees may have created stressful environments for the *E. coli* in conjunction with some

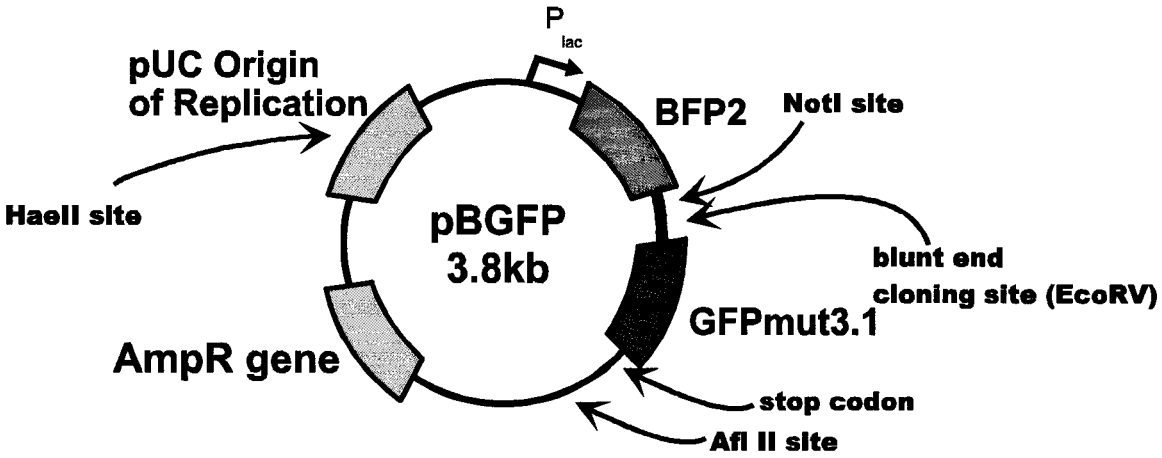
potential toxicity from the fluorescent proteins themselves. These stressful conditions may add pressure on the *E. coli* cells to recombine their plasmids. There is some evidence that growing cells at 30 degrees decreases the stress and frequency of recombination in *E. coli.*

Given the many models of RecA independent recombination, the most likely explanation of the deleted clones we observed is the replication misalignment model of Albertini et al. In this model, the template strand bunches up and the nascent strand replicates only one copy of the repeated region. The single copy that is left is actually a hybrid made up of partial copies of two templates. The sister chromosome exchange model of Lovett, Drapkin et al. results in dimerized plasmids that would be much larger than the deletion clones we observed. The replicational rearrangement model proposed by Bi, et al. could yield the observed deletion mutants. However, this model requires endonuclease and ligase activity as part of the mechanism of rearrangement.

It is difficult in these experiments to confirm that the deletion clones contained a hybrid of BFP and GFP. Out of 238 total codons, all of the differences between the two genes occurs in the first 145 codons. If the deletion product is a hybrid of BFP and GFP, the GFP portion of the hybrid would have to start after codon 145. Because of the seamless transition from the deletion gene product to the stop codon and subsequent sequence that existed after GFP, it is easy to imagine that a hybrid of BFP and GFP was created. Further experiments introducing silent mutations throughout either the BFP or GFP genes could confirm the creation of hybrid genes. The deletion products would contain some unique sequence elements from both BFP and GFP

throughout the new hybrid gene. It would also be possible to pinpoint the approximate location where BFP transitions into GFP.

This chapter serves as a warning to researchers creating unique bacterial vectors with repeated sequences. According to the models for RecA independent recombination, even short sequences can recombine in any strain of *E. coli*. In the scenario of having two similar gene products, it is possible to mutate one of the gene products throughout its entire length with silent mutations. There are many commercial gene construction services that can produce long gene sequences from a given nucleotide sequence. This may significantly decrease the frequency of recombination between the two genes. In this work, we actually substituted one of the GFP mutants with a DsRed fluorescent protein that has little sequence similarity with GFP. There has been little evidence of rearrangements with the pGRFP series of vectors.




A

B

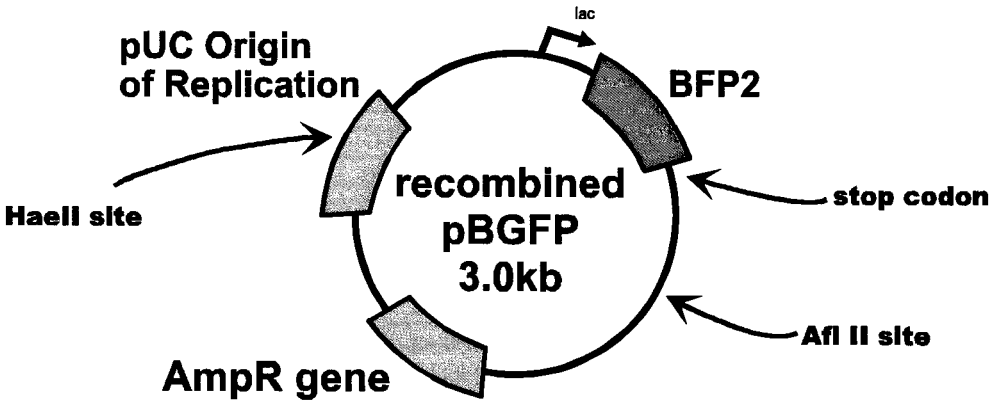



**Figure 43. Comparison of pBGFP before and after recombination.** A) Layout of native pBGFP vector. The Not I, Afl II, and Hae II restriction sites are shown for reference. B) Layout of recombined pBGFP with deleted GFP gene. The region after the BFP gene is identical to the region after the GFP gene in (A).

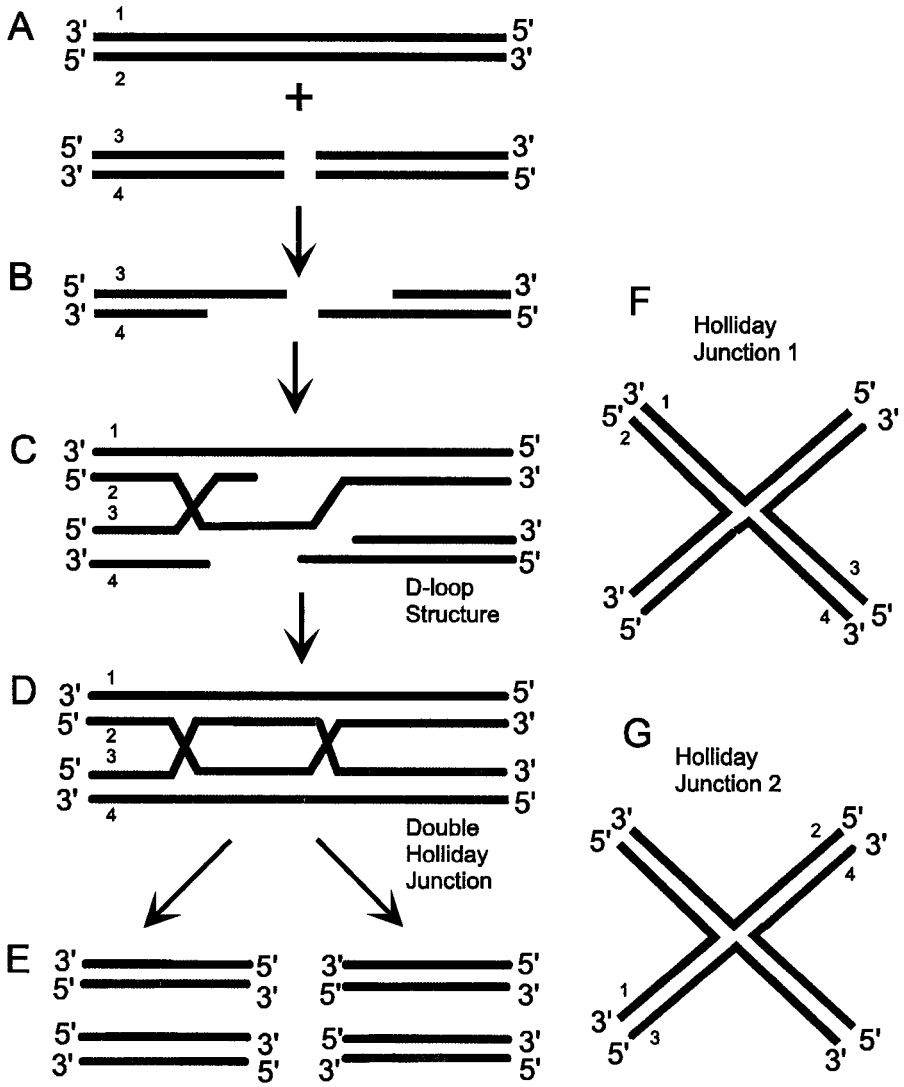



**Figure 44. Mechanism of RecA-dependent recombination.** A) RecA-dependent recombination begins with a double stranded break. Here it is shown on strands 3 and 4. B) RecBCD unwinds and chews the strands back until a $\chi$ site is reached. This process leaves 3' overhangs. C) RecA binds to the 3' overhangs and causes strand intercalation into strands 1 and 2, forming a D-loop Structure. D) Replication and ligation fills in gaps leading to the Double Holliday Junction. The migration of the Double Holliday Junction is not shown. E) Resolvase (RuvC) breaks the Holliday Junctions in two possible directions each to yield the two possible sets of products. F and G) The two Holliday Junctions are shown in a slightly different way. It easier to see how RuvC cuts the junctions from this view. Note that the strand number labels are always placed on the same end of each strand from A-G to facilitate correlations between A-E and F-G views.

Figures adapted from Kowalczykowski 2000 and
http://www.oup.co.uk/images/best.textbooks/genesvii/gifs/new1404.GIF .

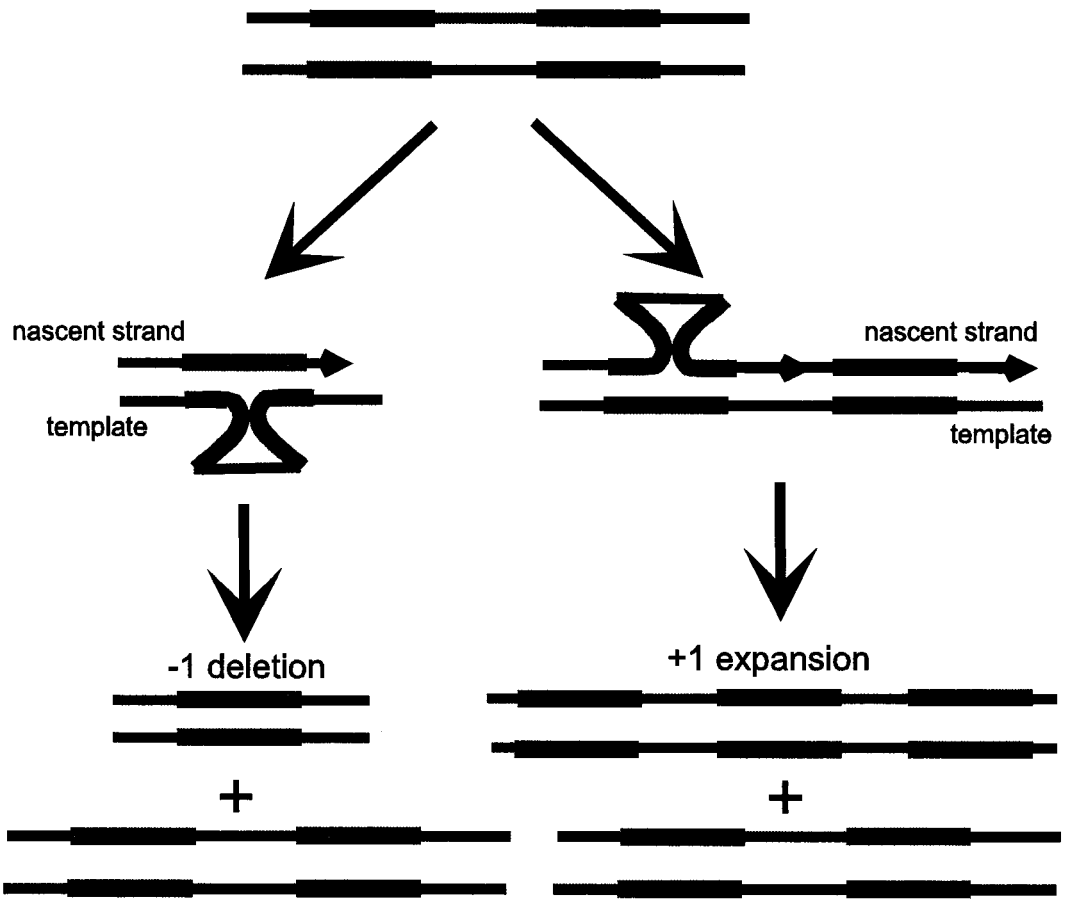



**Figure 45. Replication misalignment ("slippage") model.** Misalignment of the nascent strand can lead to either the deletion of one of the repeats or the expansion to three repeats. In the deletion scenario, the template strand bunches up on the growing nascent strand during replication. In the expansion scenario, the nascent strand bunches up after replication of both repeats. This allows the second repeat to be replicated again.

Figure adapted from Bzymek, et al. 2001.

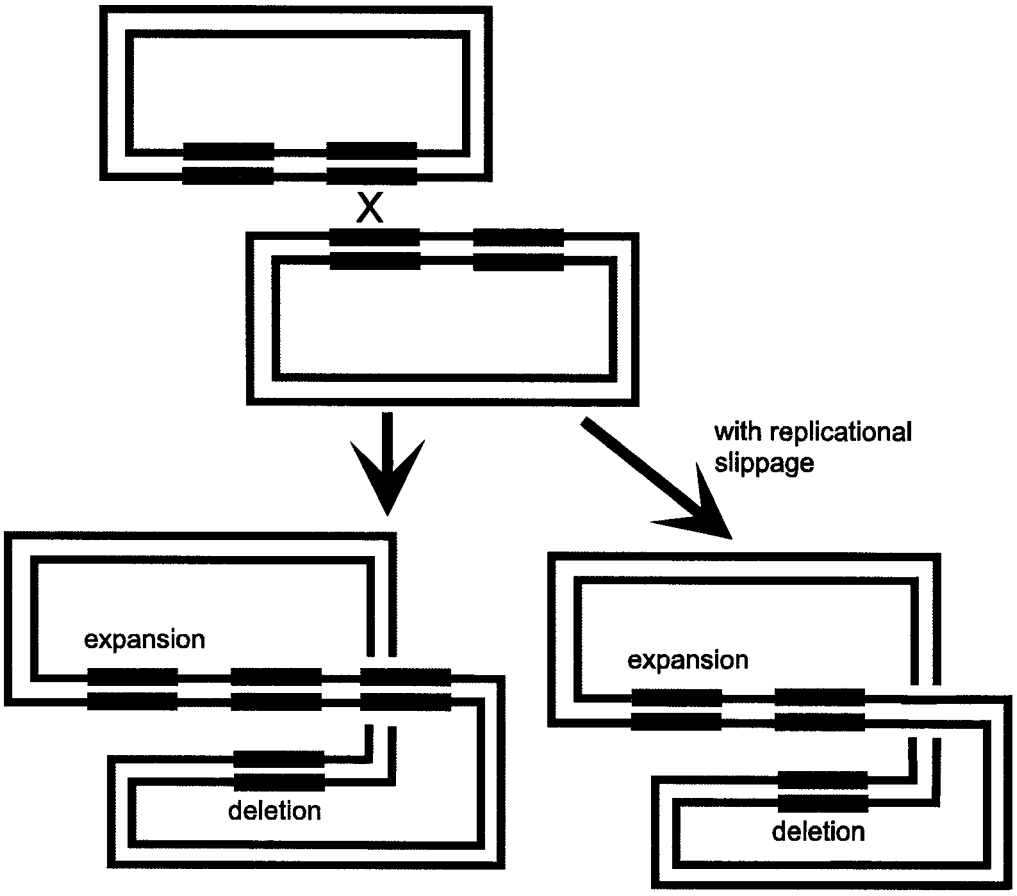



**Figure 46. Sister chromosome exchange model.** If recombination occurs between two sister strands at a replicational fork, a deletion / triplet repeat dimer product will be formed. If slippage also occurs during this process, a deletion / duplication dimmer can be formed.

Figures adapted from Byzmek and Lovett 2001.

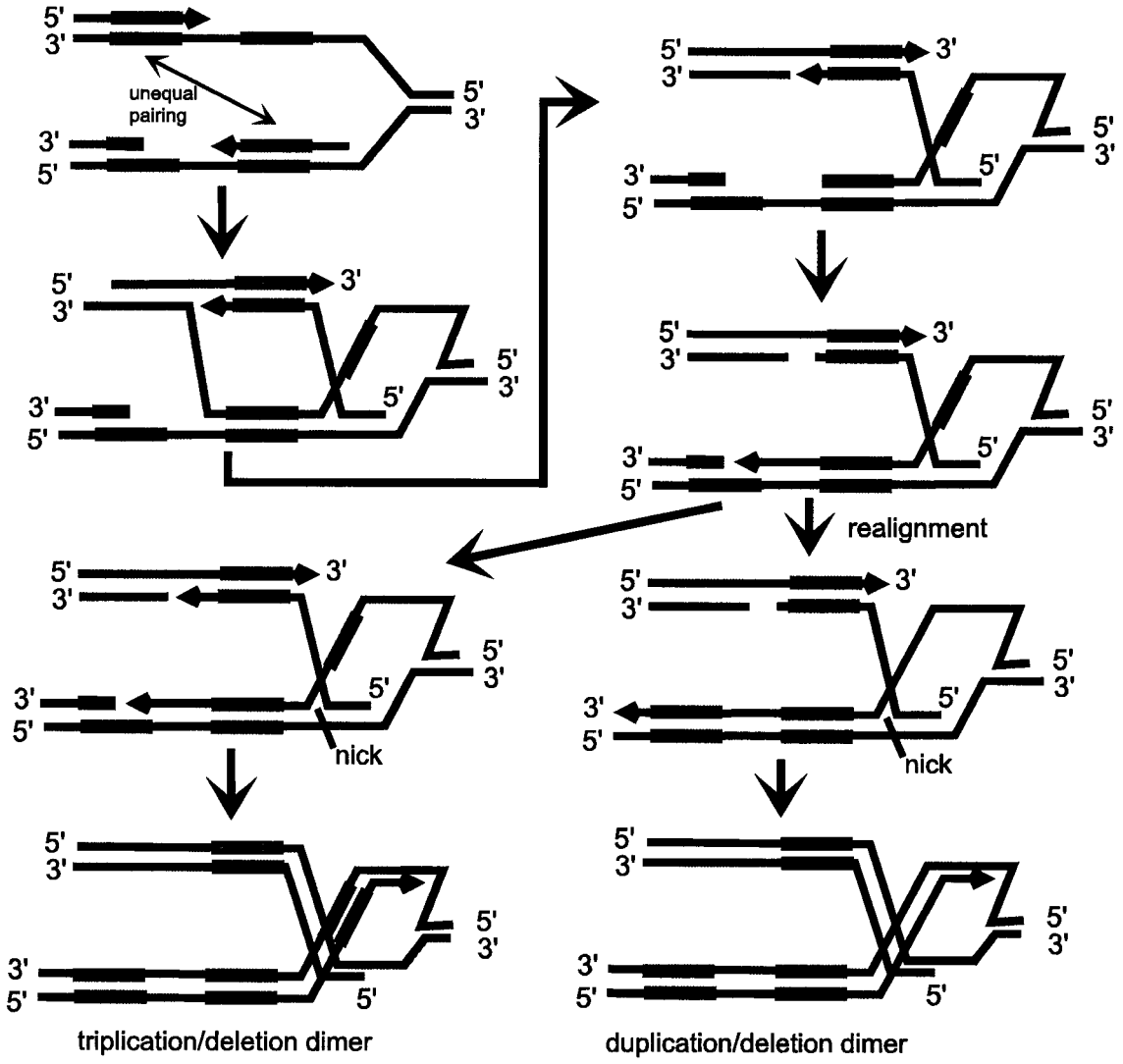



**Figure 47. Sister chromosome exchange model in depth.** The same process illustrated in figure 9 is now shown in more depth at a replicational fork. Sister strands pair during replication. After nicking and synthesis, we typically end up with one triplication and one deletion. Since the starting molecule is circular, the final product is a dimerized plasmid containing both the triplication and deletion on opposite ends. If realignment occurs, a duplication/deletion dimer results.

Figures adapted from Byzmek and Lovett 2001.

# *Chapter 7. Conclusion – A Glimpse of the Future*

## *A) Single Cell Whole Genome Amplification*

(In collaboration with Monica Orellana, Institute for Systems Biology, Seattle, WA)

If one looks at the microorganisms that have been largely studied in biology, they tend to be organisms that can easily be cultured. By culturing, one obtains many identical copies of the original organism. Even in studying the biology of eukaryotes, cell lines are commonly used, because it is possible to keep regenerating copies of the same cells. This feature is vitally important in biology to control for countless confounding factors present in the diversity of cells. In terms of genetics, it has traditionally been very difficult to know anything about the genetic content of microorganisms if they cannot be cultured.

In fact, the vast majority of microorganisms in the environment were not even known to exist twenty years ago. That is because a great majority of microorganisms are unculturable using known methods. The previous estimate was that bacterial and archael species numbered in the thousands. We now know that there are probably several million species.[1] This estimate has increased dramatically due to the use of PCR to amplify ribosomal RNA genes.[2]

---

[1] Torsvik, V., L. Ovreas, and T.F. Thingstad. (2002). "Prokaryotic diversity: magnitude, dynamics, and controlling factors." Science **296**: 1064-1066.

[2] Head, I.M., J.R. Saunders, and R.W. Pickup. (1998). "Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms." Microb. Ecol. **35**: 1-21.

The widespread use of rRNA gene amplification has allowed us to identify many new species. However, there is still not much known about what kind of organisms many of these new species are. It has been impossible to do biochemical tests or to sequence large amounts of genomic DNA from microorganisms that cannot be cultured. More recently, the concept of a "metagenome" has been presented.[3] There are several groups that are in the process of harvesting a large number of cells from environmental soil or marine sources. Then, these cells are lysed and harvested genomic DNA is subcloned into BACs to create large libraries. Then, these libraries can be scanned looking for new 16S rRNA sequences. When a rRNA gene of interest is found, the rest of the BAC can be sequenced to discover what other genes are present in that organism. Edward Delong's group recently used a similar method to discover a type of rhodopsin in a species of bacterium that was only thought to exist in archael species.[4]

The segregation of environmental genomic DNA into BAC sized fragments is a great advance over the simple PCR amplification of just the region around rRNA genes. However, it may be possible in the near future to create entire replicas of genomes from a single cell of that species. In this experiment, we describe the use of flow cytometers to isolate individual eukaryotic phytoplankton cells. Then, these cells are lysed, and their genomes are amplified by rolling circle amplification. Whole genome amplification by RCA has been demonstrated in the past from as little as 1-

---

[3] Rondon, M.R., P.R. August, A.D. Bettermann et al. (2000). "Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms." Applied and Environmental Microbiology **66**: 2541-2547.

[4] Beja, O., L. Aravind, E.V. Koonin et al. (2000). "Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea." Science **289**: 1902-1906.

10 human cells.[5] In general, RCA has been shown to do a good job of evenly representing genomic sequence with minimal bias. Amersham Biosciences recently began marketing Genomiphi, an RCA kit based on their ϕ29 DNA polymerase that is optimized for genomic amplification. The final product of the RCA reactions are a representation of the contents of the cell's genome. We should be able to PCR amplify genes from the RCA product. Alternatively, the RCA products could be sheared and subcloned into plasmid vectors. This would allow us to partially or completely sequence the genome.

For our experiment, we flow sorted cultured *Chlorella vulgaris* from a homogenous pool. *Chlorella* are eukaryotic phytoplankton that exist in abundance in the oceans. We used cultured cells so that we could reliably amplify genes that are known to be present. However, the same flow sorting techniques can very well be used on complex cell populations from any environmental source. Many groups typically analyze filtered sea water or soil samples by flow cytometry to assess scattering properties and fluorescence characteristics of cells.

*Chlorella vulgaris* cells were obtained from culture in artificial sea water. Sets of *Chlorella vulgaris* cells were flow sorted into standard PCR tubes containing 9 μl of Genomiphi Sample Buffer. Two tubes contained a single cell, two tubes contained 10 cells, two tubes contained 100 cells, and two tubes contained 1000 cells. The buffer was frozen in a dry ice ethanol batch and then heated to 95 degrees C in a thermocycler two times to lyse the cells. The buffer was then heated continuously for

[5] Dean, F.B., S. Hosono, L. Fang et al. (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc. Natl. Acad. Sci. **99**: 5261-5266.

4 minutes at 95 degrees C to denature genomic DNA so that the random hexamer primers in the Genomiphi buffer can anneal. 9 µl of Reaction Buffer and 1 µl of Enzyme Mix were added according to the recommended Genomiphi protocol. The reactions were then incubated for 30 degrees for 24 hours and heated to 65 degrees for 10 minutes to inactivate the polymerase.

PCR reactions were performed using Biolase Diamond DNA Polymerase (Bioline) according to the recommended protocol in 20 µl reaction sizes. Primers were designed within the rubisco large subunit (rbcL) gene that is present in *Chlorella vulgaris*' chloroplast DNA. Rubisco is one of the primary genes that is needed for photosynthesis. The primers were designed to be approximately 340bp apart. They were added at 0.4µM final concentration each. 1 µl of each Genomiphi product was used as template for each PCR reaction. Thermocycling was done using touchdown PCR techniques with an annealing temperature that started at 59°C and finished at 55°C.

The PCR products are shown in figure 48. One half of the Genomiphi reactions from one or ten cells yielded a product. All of the Genomiphi reactions from 100 or 1000 cells yielded a product. This shows that we are able to non-specifically amplify genomic DNA from a single cell in some cases. We did not specify for the Genomiphi reactions which portions of the genome we wanted to amplify. Yet, the Genomiphi reactions contained the rbcL gene that we were able to PCR amplify. The rbcL gene was almost certainly amplified by the Genomiphi reaction from chloroplast DNA contained within the sorted cells. It is highly unlikely that chloroplast DNA

from the surrounding culture medium was amplified, since we obtained negative PCR results in some cases even when up to ten cells were sorted. When larger numbers of cells were used, consistently positive PCR products were obtained.

It would be interesting to PCR amplify other genes from these same Genomiphi reaction products. In this study, we amplified a section from rbcL, which is present on the chloroplast DNA. Because the chloroplast DNA exists in thousands of copies per cell, we predict that the chloroplast DNA may have been preferentially amplified. It would be interesting to see if genes on *Chlorella*'s chromosomal DNA can also be amplified using this technique.

In similar fashion to the bacterial work documented in this thesis, one could imagine a high throughput process that uses flow cytometers to sort single cells from a complex environmental sample. These cells could be lysed, and their genomic contents could be amplified with RCA. Then, markers such as the rRNA genes could be PCR amplified from each RCA product. If a new or interesting rRNA signature is seen, the RCA product could be used as the starting point for further investigation. In the not too distant future, one can imagine that entire genomes could be quickly and easily sequenced from any new and interesting organisms found.

This simple experiment was designed to show the potential for single cell genomic amplification. We used a cell sorter in this study to separate out individual *Chlorella* cells from a dense culture. In some cases, single cells were sorted into tubes. Then, genomic DNA was amplified in a non-specific manner. We then showed that we could interrogate the amplified DNA for individual gene sequences. This is

the ultimate generalization of the clone isolation work presented in the main body of this thesis. Instead of bacterial cells containing a plasmid with an insert of interest, we are sorting whole organisms with large amounts of genomic content. We can generalize the term "clone" to indicate any type of individual cell that differs genetically in some way compared with surrounding cells. One can even imagine repeating this experiment using cancer cells to look at the types and numbers of mutations at multiple loci in a single cell.

### *B) Concluding Remarks*

In some sense, this experiment involving *Chlorella* represents an extreme example of what we have achieved in this work with *E. coli*. With *Chlorella*, the packets of genetic material that we are interested in are very large, since we are interested potentially in sequences throughout the genome. However, the criteria that we use to select cells for sorting is relatively simple. We are typically looking for physical properties such as size or maybe some cellular fluorescence.

In all of the projects involving *E. coli*, the genetic material of interest is very small and very well defined. Because we are cloning inserts into plasmid vectors, the genetic material we are interested in is typically smaller than 5kb and is located between two well-defined markers on each plasmid vector. However, the criteria for sorting becomes very complex, because we only want to sort cells that have a well-defined genetic property. In the "Sequencing Project" chapter and "Protein Mutagenesis Project" chapter, we are primarily interested in whether an insert is present within a plasmid vector. In the "Novel Method for Detecting DNA-Protein

Interactions" chapter we are interested in whether a particular cloned DNA fragment interacts with a co-expressed protein. It is this translation from a biological property into a readable physical property that is really at the heart of this work. We use plasmid vectors that encode internal biological states as a fluorescent signal that is visible from the outside.

The bacteria in this work can be thought of as individual envelopes that carry user specified information encoded as DNA. These envelopes can also be coded with fluorescent markings to give an indication for its contents. The plasmid vectors we have designed are the biological equivalent of an algorithm that takes an envelope's contents and decides what fluorescent markings to use. The flow cytometer acts as an envelope sorting machine that can decide whether to keep or discard envelopes. The flow cytometer can also separate incoming envelopes into individual pieces for further study.

In the most simple case, one fluorescent marking should have been sufficient to indicate whether a particular envelope should be kept or discarded. If the fluorescent marking is present, keep the envelope. If the fluorescent marking is not present, discard the envelope. This was what was attempted in the first vector we designed, pSE-GFP+. However, it turns out that *E. coli* have too much variability in terms of how much fluorescent signal they produce. Therefore, we moved to the two fluorescent protein model. In this model, one fluorescent protein serves as the internal control for the other. It is as if the intensity of the fluorescent markings on the envelopes is unreliable. Therefore, the flow cytometer compares the intensity of two

fluorescent markings. If one fluorescent marking is brighter than the other, then the envelope should be kept.

In this manner, we have designed plasmid vectors that act as fluorescence encoding "algorithms". The pGRFP series of vectors have proven their reliability in indicating to flow cytometers whether an insert is present or not. In chapters 2 and 3, we describe the use of this vector in the context of a shotgun sequencing project as well as to characterize a large number of mutants of a human enzyme. For unrelated technical reasons, we were unable to use the flow cytometer for the mutant sequencing project. However, we have proven our ability to successfully differentiate between single bacteria based on the presence or absence of insert.

In chapter 4, we describe the creation of the pGFPpDsRed series of vectors that can successfully encode DNA-protein interactions as fluorescent signals. Again, this is another type of algorithm that encodes a different biological state as a physical property of individual *E. coli* cells. This algorithm is somewhat unique, because it uses the principal of transcription repression caused by a bound protein. Although repressor proteins have been known to work on this principle, this is one of the first uses of repression as an actual assay for binding. It will be interesting to see if this principle works with other proteins and binding sites.

In the theoretical chapters, we attempt to explain some of the stranger artifacts that we have observed. In chapter 5, we attempt to explain our observation that bacteria do not develop fluorescence until late log to stationary phase. In a sense, this is a breakdown of our abstraction. If envelopes are to be marked with fluorescent

signals, why can we only see the signals at a time much later than we expect? This is one of the perils of working with complex biological systems.

In chapter 6, we describe our observations involving recombination with one of the earlier plasmid prototypes. Again, this is a breakdown of our abstraction. Our algorithm is stipulating that a particular fluorescent signal be marked on an envelope, but through recombination, the algorithm sometimes loses its ability to create one of the fluorescent colors. This is another example of the dangers of oversimplifying complex biological systems. There will always be some situations in which the system does not follow the model.

We have shown that the algorithms effectively code envelopes with fluorescent markings. Furthermore, we have used flow cytometers to separate a stream of envelopes into individual units. The next question is how to effectively process the individual envelopes once they are separated. With the advancement of single cell amplification techniques, we may be able to reliably amplify plasmid DNA from single sorted cells. This would obviate the need for culturing and would make the issue of cell viability irrelevant. Advances in liquid handling and carrier format similar to the linear tape technologies presented in the "Introduction and General Background" chapter would streamline the processing of DNA amplification and sequencing reactions in batches of tens of thousands.

Looking further, microfabricated DNA chips and microfluidics will allow for the fast processing of thousands of samples in a very small space. Several groups are developing a high level of proficiency at moving small volumes of fluids around on a

microfabricated chip.[6] One can imagine a chip in which the flow cytometer sorts onto one spot on a chip. After each cell is sorted, the chip transports transports the cell into its own reaction chamber. This would probably occur quickly at rates over 10 cells per second. Then, cell lysis and DNA amplification can occur in each individual chamber. Even sequencing reactions could proceed completely on the chip. The Mathies group has developed microfabricated chips that have the capability to thermocycle the contents of small chambers. They have demonstrated the ability to perform sequencing reactions and to run the sequencing reaction products in small capillary channels etched into the chip.[7]

Our work lays the foundation for high throughput processes based on flow cytometers. Coupled with the right technologies, our methods have the potential to isolate hundreds of thousands of clones per day, assess mutability of proteins quickly and efficiently, or solve regulatory networks within large genomes. With this work, flow cytometers have truly become a tool for answering scientific questions in the genomic age.

---

[6] Thorsen, T., S.J. Maerkl, S.R. Quake. (2002). "Microfluidic Large Scale Integration." Science **298**: 580-584.

[7] Paegel, B.M., R.G. Blazej, R.A. Mathies. (2003) "Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis." Curr. Opin. Biotechn. **14**: 42-50.

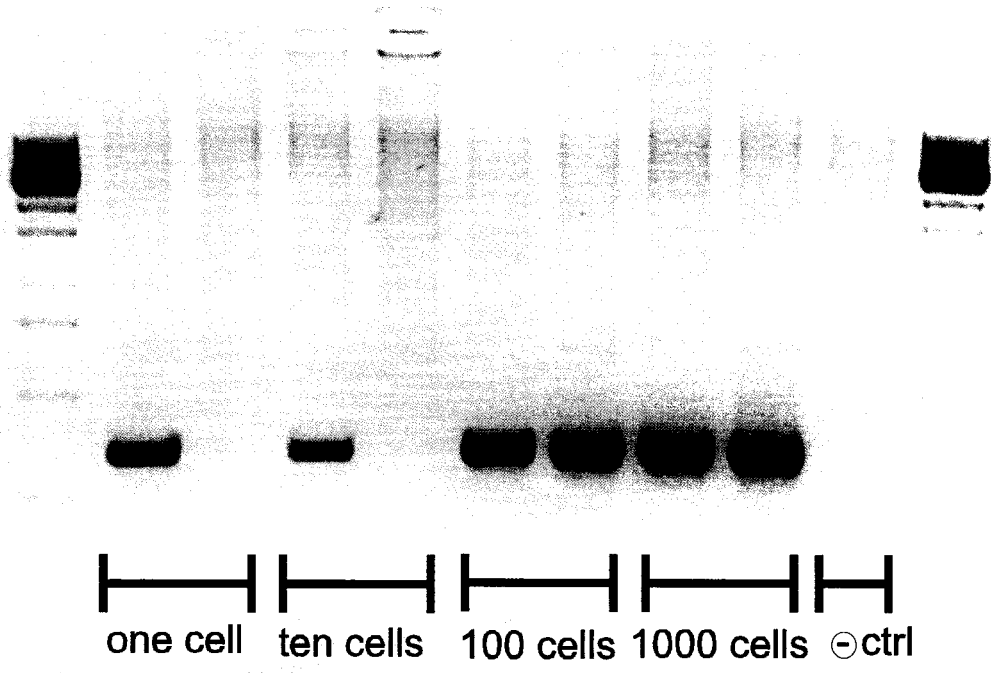



**Figure 48. Gel showing PCR amplification from Genomiphi product.** A 350bp portion of the rbcL gene present on the chloroplast of *Chlorella vulgaris* was PCR amplified from Genomiphi amplified DNA from small numbers of sorted cells. The number of original *Chlorella* cells used is marked for each lane. The negative control PCR reaction used Genomiphi product that had no *Chlorella* cells added.

**List of References**

Adams, M.D., S.E. Celniker, et al. (2000). "The genome sequence of Drosophila melanogaster." Science **287**: 2185-2195.

Albertini, A.M., M. Hofer, M.P. Calos, J.H. Miller. (1982). "On the Formation of Spontaneous Deletions: the Importance of Short Sequence Homologies in the Generation of Large Deletions." Cell **29**:319-328.

Allgood, N.D. and T.J. Silhavy. (1991). "*Escherichia coli* xonA (sbcB) Mutants Enhance Illegitimate Recombination." Genetics **127**: 671-680.

Asai, T., D.B. Bates, and T. Kogoma. (1995). "DNA Replication Triggered by Double-Stranded Breaks in *E. coli*: Dependence on Homologous Recombination Functions." Cell **78**: 1051-1061.

Baird, G.S., D.A. Zacharias, and R.Y. Tsien. (2000). "Biochemistry, Mutagenesis and Oligomerization of DsRed, a Red Fluorescent Protein from Coral." Proc. Natl. Acad. Sci. **97**: 11984-11989.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). "Determinants of a protein fold. Unique features of the globin amino acid sequences." J Mol Biol **196**: 199-216.

Beasley, J.R. and M.H. Hecht. (1997). "Protein Design: The Choice of de Novo Sequences." J. Biol. Chemistry **272**:2031-2034.

Beja, O., L. Aravind, E.V. Koonin et al. (2000). "Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea." Science **289**: 1902-1906.

Bernard, L., C. Courties, et al. (2001). "A new approach to determine the genetic diversity of viable and active bacteria in aquatic ecosystems." Cytometry **43**:314-321.

Bertram, S., F.T. Hufert, et al. (1995). "Detection of DNA in single cells using an automated cell deposition unit and PCR." Biotechniques **19**:616-620.

Bertrand-Burggraf, E., S. Hurstel, M. Duane, and M. Schnarr. (1987). "Promoter properties and negative regulation of the uvrA gene by the LexA repressor and its amino-terminal DNA binding domain." J. Mol. Biol. **193**: 293-302.

Bevis, B.J. and B.S. Glick. 2002. Rapidly maturing variants of the Discosoma red fluorescent protein (DsRed). Nature Biotechnology. 20:83-87.

Bi, X. and L.F. Liu. (1996). "A Replicational Model for DNA Recombination between Direct Repeats." J. Mol. Biol. **256**: 849-858.

Bi, X. and L.F. Liu. (1994). "RecA-Independent and RecA-dependent Intramolecular Plasmid Recombination" J. Mol. Biol. **256**: 849-858.

Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). "Deciphering the message in protein sequences: tolerance to amino acid substitutions." Science **247**: 1306-10.

Bzymek, B. and S. Lovett. (2001). "Instability of Repetitive DNA Sequences: The Role of Replication in Multiple Mechanisms." Proc. Natl. Acad. Sci. **98**: 8319-8325.

The C. elegans Sequencing Consortium. (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science **282**: 2012-2018.

The Celera Genomics Sequencing Team. (2001). "The Sequence of the Human Genome." Science **291**: 1304-1351.

Chen, H., M. Bjerknes, R. Kumar, and E. Jay. 1994. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. Nucleic Acids Research. 22: 4953-4957.

Chong, S.S., E. Almqvist, et al. (1997). "Contribution of DNA sequence and CAG size to mutation frequencies of intermediate alleles for Huntington disease: evidence from single sperm analyses." Human Molecular Genetics **6**: 301-309.

Cline, J. and H. Hogrefe. (2000). "Randomize Gene Sequences with New PCR Mutagenesis Kit." Stratagies **13**:157-161.

Clontech Laboratories. "Diversify PCR Random Mutagenesis Kit." Clontechniques **October 1999**: 14-15.

Cody, C.W., D.C. Prasher, et al. (1993). "Chemical Structure of the Hexapeptide Chromophore of the Aequorea Green-Fluorescent Protein." Biochemistry **32**: 1212-1218.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. (1997). "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale." Science **278:** 680-686.

Dean, F.B., S. Hosono, L. Fang et al. (2002). "Comprehensive human genome amplification using multiple displacement amplification." Proc. Natl. Acad. Sci. **99**: 5261-5266.

Dolan, J.W., C. Kirkman, and S. Fields. (1989). "The yeast STE12 protein binds to the DNA sequence mediating pheromone induction." Biochemistry **86**: 5703-5707.

Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998). "Rates of spontaneous mutation." Genetics **148**:1667-86.

Edelman, G. M. & Gally, J. A. (2001). "Degeneracy and complexity in biological systems." Proc Natl Acad Sci **98:** 13763-8.

Elowitz, M.B., A.J. Levine, E.D. Siggia, and P.S. Swain. (2002). "Stochastic Gene Expression in a Single Cell." Science **297**: 1183-1186.

Encell, L.P., D.M. Landis, and L.A. Loeb. (1999). "Improving enzymes for cancer gene therapy." Nat. Biotechnol. **17**:143-147.

Ewing B., L. Hillier, M. Wendl, P. Green. 1998. Basecalling of automated sequencer traces using phred. Genome Research. 8: 175-194.

Fields, S. and O. Song. (1989). "A Novel Genetic System to Detect Protein-Protein Interactions." Nature **340**: 245-246.

Friedberg, E.C. et al. (1995) DNA Repair and Mutagenesis. Washington DC: ASM Press.

Fromant, M., S. Blanquet, and P. Plateau. (1995). "Direct Random Mutagenesis of Gene-sized DNA Fragments using Polymerase Chain Reaction." Anal. Biochem. **224**:347-353.

Gordon, D., C. Abajian, and P. Green. 1998. Consed: A graphical tool for sequence finishing. Genome Research. 8:195-202.

Gross, L.A., G.S. Baird, et al. (2000). "The Structure of the Chromophore within DsRed, a Red Fluorescent Protein from Coral." Proc. Natl. Acad. Sci. **97**:11990-11995.

Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**: 994-9.

Hanahan, D., and Weinberg, R.A. (2000). "The hallmarks of cancer". Cell **100**:57-70.

Harris, R.S., Bishop, K.N., Sheehy, A.M., Craig, H.M., Petersen-Mahrt, S.K., Watt, I.N., Neuberger, M.S., and Malim, M.H. (2003). "DNA deamination mediates innate immunity to retroviral infection." Cell **113**:803-809.

Head, I.M., J.R. Saunders, and R.W. Pickup. (1998). "Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms." Microb. Ecol. **35**: 1-21.

Inouye, S., H. Ogawa, et al. (1997). "A Bacterial Cloning Vector using a Mutated Aequorea Green Fluorescent Protein as an Indicator." Gene **189**: 159-162.

The International Human Genome Mapping Consortium. (2001). "A Physical Map of the Human Genome." Nature **409**: 934-941.

Iyer, V.R., C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, P.O. Brown. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature **409**: 533-538.

Kalakonda, N., D.G. Rothwell DG, et al. (2001). "Detection of N-Ras codon 61 Mutations in Subpopulations of Tumor Cells in Multiple Myeloma at Presentation." Blood **98**:1555-1560.

Kamtekar, S., J.M. Schiffer, J.M. Xiong et al. (1993). "Protein design by binary patterning of polar and nonpolar amino acids." Science **262**:1680-1685.

Kato, M., D.N. Frick, J. Lee, S. Tabor, C.C. Richardson, T. Ellenberger. 2001. "A complex of the bacteriophage T7 primase-helicase and DNA polymerase directs primer utilization." J. Biol. Chem. **276**:21809-20.

Kawate, H., D.M. Landis, and L.A. Loeb. 2002. "Distribution of mutations in human thymidylate synthase yielding resistance to 5-fluorodeoxyuridine". J. Biol. Chem. **277**: 36304-36311.

Kowalezykowski, S.C. (2000) "Initiation of Genetic Recombination and Recombination-Dependent Replication." Trends in Biochem. Sci. **25**: 156-165.

Kowalczykowski, S.C. and A.K. Eggleston. (1994). "Homologous Pairing and DNA Strand-Exchange Proteins." Annu. Rev. Biochem. **63**: 991-1043.

Kronstad, J.W., J.A. Holly, and V.L. MacKay. (1987). "A yeast operator overlaps an upstream activation site." Cell **50**: 369-377.

Labahn, J., O.D. Scharer, et al. (1996). " Structural Basis for the Excision Repair of Alkylation-Damaged DNA." Cell **86**: 321-329.

Lau, A., O. Scharer, et al. (1998). "Crystal Structure of a Human Alkylbase-DNA Repair Enzyme Complexed to DNA: Mechanism for Nucleotide Flipping and Base Excision." Cell **95**:249-258.

Lau, A. Y., Wyatt, M. D., Glassner, B. J., Samson, L. D. & Ellenberger, T. (2000). "Molecular basis for discriminating between normal and damaged bases by the human alkyladenine glycosylase, AAG." Proc Natl Acad Sci **97**: 13573-8.

Lee, J. and A. Goldfarb. (1991). "*lac* repressor acts by modifying the initial transcribing complex so that it cannot leave the promoter." Cell **66**:793-798.

Lee, T.I., N.J. Rinaldi, F. Robert et al. (2002). "Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*." Science **298**: 799-804.

Lieb, J.D., X. Liu, D. Botstein, P.O. Brown. "Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association." Nature Genetics **28**: 327-334.

Lizardi, P.M., X. Huang, Z. Zhu et al. "Mutation detection and single-molecule counting using isothermal rolling-circle amplification." Nature Genetics **19**: 225-232.

Lockhart, D.J., H. Dong et al. (1996). "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays." Nat Biotechnol **14**: 1675-1680.

Loeb, L.A., Loeb, K.R., and Anderson, J.P. (2003). "Multiple mutations and cancer." Proc Natl Acad Sci U S A **100**:776-781.

Lopez, P.J., J. Guillerez, R. Sousa, and M. Dreyfus. (1998). "On the Mechanism of Inhibition of Phage T7 RNA Polymerase by *lac* Repressor." J. Mol. Biol. **276**:861-875.

Lovett, S.T., P.T. Drapkin, et al. (1993). "A Sister-Strand Exchange Mechansim for recA-Independent Deletion of Repeated DNA Sequences in *Escherichia coli*." Genetics **135**: 631-642.

Lovett, S.T., T.J. Gluckman, et al. (1994). "Recombination between Repeats in *Escherichia coli* by a recA-Independent, Proximity-Sensitive Mechanism."

Low, B. (1968). "Formation of merodiploids in matings with a class of Rec- recipient strains of *Escherichia coli* K12." Proc. Natl. Acad. Sci. **60**:160-167.

Lutz, R., T. Lozinski, T. Ellinger, and H. Bujard. "Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator." Nuc. Acids Res. **29**:3873-3881.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). "Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts". Nature **424**:99-103.

Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. (1994). "Genetic studies of the lac repressor. XIV. Analysis of 4000 altered *Escherichia coli* lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence." J Mol Biol **240**:421-33.

Martin, C.T., D.K. Muller, and J.E. Coleman. (1988). "Processivity in early stages of transcription by T7 RNA polymerase." Biochemistry **27**: 3966-3974.

Matz, M.V., A.F. Fradkov, et al. (1999). " Fluorescent Proteins from Nonbioluminescent Anthozoa species." Nature Biotech. **17**:969-973.

Maynard-Smith, J. (1970). "Natural Selection and the Concept of a Protein Space." Nature **225**: 563-564.

McEntee K. (1977). "Protein X is the product of the recA gene of *Escherichia coli*." Proc. Natl. Acad. Sci. **74**: 5275-5279.

Memisoglu, A. and L. Samson. (1996). "DNA Repair Functions in Heterologous Cells." Crit. Rev. Biochem. Mol. Biol. **31**:405-447.

Miao, F., Bouziane, M., Dammann, R., Masutani, C., Hanaoka, F., Pfeifer, G. & O'Connor, T. R. (2000). "3-Methyladenine-DNA glycosylase (MPG protein) interacts with human RAD23 proteins." J Biol Chem **275**:28433-8.

Miyazaki, K. and F.H. Arnold. (1999). "Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function." J. Mol. Evol. **49:**716-20.

Nakabeppu, Y. and M. Sekiguchi. (1986). "Regulatory Mechanisms for Induction of Synthesis of Repair Enzymes in Response to Alkylating Agents: Ada Protein acts as a Transcriptional Regulator." Proc. Natl. Acad. Sci. **83**: 6297-6301.

Narlikar, G.J., H.-Y. Fan, and R.E. Kingston. (2002). "Cooperation between complexes that regulate chromatin structure and transcription." Cell **108**: 475–487.

O'Connor, T. R. (1993). "Purification and characterization of human 3-methyladenine-DNA glycosylase." Nucleic Acids Res. **21**: 5561-9.

Orphanides, G. and D. Reinberg. (2002) "A Unified Theory of Gene Expression." Cell **108**: 439-451.

Paegel, B.M., R.G. Blazej, R.A. Mathies. (2003) "Microfluidic devices for DNA sequencing: sample preparation and electrophoretic analysis." Curr. Opin. Biotechn. **14**: 42-50.

Patterson, G.H., S.M. Knobel, et al. (1997). "Use of the Green Fluorescent Protein and its Mutants in Quantitative Fluorescence Microscopy." J. Biophys **73**: 2782-2790.

Reid, B.G. and G.C. Flynn (1997). "Chromophore Formation in Green Fluorescent Protein." Biochemistry **36**: 6786-6791.

Ren, B., F. Robert, J.J. Wyrick et al. (2000). "Genome-Wide Location and Function of DNA Binding Proteins." Science **290**:2306-2309.

Roessner, C.A. and A. Ian Scott. (1995). "Fluorescnence-Based Method for Selection of Recombinant Plasmids." Biotechniques **19**: 760-764.

Rondon, M.R., P.R. August, A.D. Bettermann et al. (2000). "Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms." Applied and Environmental Microbiology **66**: 2541-2547.

Roy, R., Biswas, T., Hazra, T. K., Roy, G., Grabowski, D. T., Izumi, T., Srinivasan, G. & Mitra, S. (1998). "Specific interaction of wild-type and truncated mouse N-methylpurine-DNA glycosylase with ethenoadenine-containing DNA." Biochemistry **37**:580-9.

Samson, L. B. Derfler, et al. (1991). "Cloning and Characterization of a 3-Methyladenine DNA Glycosylase cDNA from Human Cells whose Gene Maps to Chromosome 16." Proc. Natl. Acad. Sci **88**: 9127-9131.

Sasmor, H. and J. Betz. (1990). "Symmetric *lac* operator derivatives: Effects of Half-operator Sequence and Spacing on Repressor Affinity." Gene **89**:1-6.

Shevell, D.E., B.M. Friedman, and G.C. Walker. (1990). "Resistance to Alkylation damage in *Escherichia coli*: Role of the Ada Protein in Induction of the Adaptive Response." Mutat. Res. **233**:52-72.

Schlax, P.J., M.W. Capp, and M.T. Record. (1995). "Inhibition of transcription initiation by *lac* repressor." Nucl. Acids Res. **6**:111-137.

Simon, I., J. Barnett J, N. Hannett N et al. (2001). "Serial regulation of transcriptional regulators in the yeast cell cycle." Cell **106**: 697-708.

Skandalis, A., L.P. Encell, and L.A. Loeb. (1997). "Creating Novel Enzymes by Applied Molecular Evolution." Chemistry & Biology **4**: 889-898.

Straney, S.B. and D.M. Crothers. (1987). "*lac* repressor is a transient gene-activating protein." Cell **51**:699-707.

Tang, G.L., Y.F. Wang, J.S. Bao, and H.B. Chen. 1999. Overexpression in *Escherichia coli* and Characterization of the Chloroplast Triosephosphate Isomerase from Spinach. Protein Expression and Purification. 16:432-439.

Teo, I., B. Sedgwick, et al. (1986). "The Intracellular Signal for Induction of Resistance to Alkylating agents in *E. coli*." Cell **45**: 315-324.

Thorsen, T., S.J. Maerkl, S.R. Quake. (2002). "Microfluidic Large Scale Integration." Science **298**: 580-584.

Torsvik, V., L. Ovreas, and T.F. Thingstad. (2002). "Prokaryotic diversity: magnitude, dynamics, and controlling factors." Science **296**: 1064-1066.

Tupler, R., G. Perini, and M.R. Green. (2001). "Expressing the human genome." Nature **409**: 832–833.

Vellanoweth, R.L. and J.C. Rabinowitz. 1992. The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli in vivo.* Molecular Microbiology. 6: 1105-1114.

Volkert, M.R., D.C. Nguyen, and K.C. Beard. (1986). "*Escherichia coli* Gene Induction by Alkylation Treatment." Genetics **112**: 11-26.

West, M.W. and M.H. Hecht. (1995). "Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins." Protein Sci. **4**:2032-2039.

Whoriskey, S.K., M.A. Schofield, and J.H. Miller. (1991). "Isolation and characterization of *Escherichia coli* mutants with altered rates of deletion formation." Genetics **127**:21-30.

Wyatt, M.D., J.M. Allan et al. (1999). "3-Methyladenine DNA Glycosylases: Structure, Function, and Biological Importance." Bioessays **21**:668-676.

Wyrick, J.J., J.G. Aparicio, T. Chen et al. (2001). "Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins." Science **294**: 2357-2360.

Xiong, H., B.L. Buckwalter, H.-M. Shieh et al. (1995). "Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides." Proc. Natl. Acad. Sci. **92**:6349-6353.

Yamada, M., M. Kubo, et al. (1991). "Promoter Sequence Analysis in *Bacillus* and *Escherichia*: Construction of Strong Promoters in *E. coli*." Gene **99**: 109-114.

Yamagata, Y., M. Kato, et al. (1996). "Three-Dimensional Structure of a DNA Repair Enzyme, 3-Methyladenine DNA Glycosylase II, from *Escherichia coli*. Cell **86**:311-319.

Yi, T.-M., D. Stearns, and B. Demple. (1988). "Illegitimate recombination in an *Escherichia coli* plasmid: modulation by DNA damage and a new bacterial gene." J. Bacteriol. **170**:2898-2903.

Zhao, R., A. Natarajan, F. Srienc. (1999)."A Flow Injection Flow Cytometry System for On-Line Monitoring of Bioreactors." Biotech. & Bioeng. **62**: 609-617.

Zuker, M., D.H. Mathews and D.H. Turner. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology. Kluwer Academic Publishers, 1999.

## Vita

Juno Choe was born in Minneapolis, MN, but has lived in New England, California, and South Korea. He currently lives in Seattle, WA while pursuing M.D. and Ph.D. degrees. He graduated from the Massachusetts Institute of Technology with a Bachelor of Science in Electrical Engineering and Computer. His current professional interests include biological research, clinical medicine, and software design and development. He is the author of Spigot software that is currently licensed for use with Influx flow cytometers manufactured by Cytopeia, Inc. In 2003, he earned a Doctor of Philosophy from the Department of Molecular Biotechnology at the University of Washington.