ILLUMINATING CLONAL DYNAMICS: DEVELOPMENT AND USE
OF A HIGH-THROUGHPUT CELLULAR BARCODING SYSTEM
TO TRACK CLONAL EVOLUTION

APPROVED BY SUPERVISORY COMMITTEE

Matthew Porteus, M.D., Ph.D.

Pier Paolo Scaglioni, M.D.

Lani Wu, Ph.D.

Alec Zhang, Ph.D.

DEDICATION

To my parents, Gran and Cathy Porter, for your many years

of unwavering love and support.

ILLUMINATING CLONAL DYNAMICS: DEVELOPMENT AND USE OF A HIGH-THROUGHPUT CELLULAR BARCODING SYSTEM TO TRACK CLONAL EVOLUTION

by

SHAINA N. PORTER

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

December, 2012

ACKNOWLEDGEMENTS

I have a large number of people I need to thank for helping me get to the finish line of the marathon that is graduate school. I would like to thank my Ph.D. mentor, Matt Porteus, for allowing me to join his group and having the faith in me to let me develop this project from the ground up. Also, I am grateful to my thesis committee members, Dr. Alec Zhang, Dr. Lani Wu, and Dr. Pier Paolo Scaglioni, for their guidance and help in the completion of this project. I also owe a big thank you to Dr. Woodring Wright of UT Southwestern for his numerous helpful ideas and suggestions. We are very grateful for your support and interest in this project. Eric Kildebeck was instrumental in getting the data handling aspects of this project off the ground by writing the first program to analyze the barcode sequencing data. David Mittelman and his lab at Virginia Tech, and specifically graduate student Gareth Highnam, have literally rescued me from drowning in the vast sea of data this project has produced. In the short time we have been collaborating, they have written the programs responsible for all of the barcode sequencing data analysis, programs able to do things that I had previously only dreamed about. I am hugely grateful and look forward to our continued collaboration.

Next, I would like to thank the three labmates who became my family away from home, the "Dallas Crew" who moved to Stanford from UT Southwestern in 2010. Richard, Josh, and Eric, you guys are the best, and I'm glad we've gotten to have so many adventures together, even though my participation had to be coerced at times. I would also like to thank the rest of the Porteus Lab members, both past and present, for their contributions to making the lab a fun and intellectually stimulating place work and

for helpful discussions. I would specifically like to express my gratitude to Shondra Pruett-Miller for her friendship and mentorship during our time together in the lab. Also, thanks to Tina Dann for teaching me how to make lentivirus, which has been a key part of my project. A big thank you goes to Stacey Wirt, the Stanford Porteus Lab rock star, for her willingness to be a listening ear, give advice, and provide expert help with experiments, particularly my mouse xenografts.

Finally, I would like to thank the numerous friends and family members who have helped me along during this grad school journey, especially my wonderfully supportive and loving parents, and my brothers Scott and Cody. I love you all.

This work would not have been possible without the funding support of an Innovation Award from the Alex's Lemonade Stand Foundation for Childhood Cancer. Alexandra Scott was an amazing young girl with a mission to raise money to find a cure for all kids with cancer even as she was fighting her own battle with neuroblastoma. I am honored that her foundation supported this project, and hope that this project and my future endeavors will contribute to reaching Alex's goal.

ILLUMINATING CLONAL DYNAMICS: DEVELOPMENT AND USE OF A HIGH-THROUGHPUT CELLULAR BARCODING SYSTEM TO TRACK CLONAL EVOLUTION

SHAINA N. PORTER, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2012

Supervising Professor: Matthew H. Porteus, M.D., Ph.D.

It is increasingly recognized that tracking the clonal dynamics of large populations is important in understanding aspects of cancer and stem cell biology. Attempts to track the contributions of individual cells to the clonality of large homogenous populations, however, have been constrained by limitations in sensitivity and complexity. We have created an efficient and high throughput method to overcome these limitations by harnessing the power of viral marking and next generation sequencing technology, allowing us to track the clonal contributions of many thousands of cells with minimal perturbations to the population as a whole. We have applied this system to several of the most commonly used cell lines in biological research in order to validate this system and gain valuable biological insight into these ubiquitous research

tools. Cell lines, often continuously passaged for years, are often assumed to be clonal, the most fit and stable clone having been selected during establishment and early passages. However, our results show that ongoing genomic and/or epigenomic instability within these cells leads to proliferative instability, resulting in the divergence of clones from one another and revealing unexpected clonal dynamics and rapid clonal evolution. These results have profound implications for the experimental use of cell lines, as well as broad applications in the fields of stem cells and cancer biology, among others.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

Ellis, B. L., Hirsch, M. L., **Porter, S. N.**, Samulski, R. J., Porteus, M. H. (2012). "Zinc-finger nuclease-mediated gene correction using single AAV vector transduction and enhancement by Food and Drug Administration-approved drugs." Gene Ther.

Pruett-Miller, S. M., Reading, D. W., **Porter, S. N.**, Porteus, M. H. (2009). "Attenuation of zinc finger nuclease toxicity by small-molecule regulation of protein levels." PLoS Genet **5**(2): e1000376.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF DEFINITIONS

BP – Base Pair

FISH – Fluorescence In Situ Hybridyzation

GFP – Green Fluorescent Protein

HSV-TK – Herpes Simplex Virus Thymidine Kinase

LTR – Long Terminal Repeat

MOI – Multiplicity of Infection

PCR – Polymerase Chain Reaction

TALEN – Transcription Activator-Like Effector Nuclease

UBC – Ubiquitin C

ZFN – Zinc Finger Nuclease

CHAPTER ONE
Introduction and Review of Literature

INTRODUCTION

Cells reproduce by replicating their DNA before splitting into two daughter cells which each receive one set of genetic information, resulting in two identical clones of the original parent cell. While DNA replication and cell division are generally very accurate processes, errors do occur, and can accumulate over time as a cell continues to divide, passing down any changes to the next generation. This means that the more cell divisions, or generations, that have occurred since the last common ancestor of two cells, the more likely those cells are to have acquired genetic or epigenetic differences that may affect cellular phenotypes.

Because of these acquired changes, cells within a homogeneous population such as a tumor, a normal tissue, or a cell line may be indistinguishable from one another on the surface, but may in fact be quite heterogeneous, mixtures of clones that may differ in a myriad of ways. Inability to detect this heterogeneity or quantify the contributions of different clones to a population may result in measurements that may not accurately reflect the status of any individual cell in that population. Nor can we detect with current techniques how many differently behaving clones are present in that population, or the dynamics of clones within that population. Lose valuable biologic information in the noise of many different clones.

Therefore, being able to track and identify clones within a heterogeneous population can provide important information about cell potential, rates of cell

proliferation and death, clonal fitness and competition within the population, the effects of genetic alterations on cell phenotype, as well as the physical location and distribution of clones. This information has important implications for the treatment for cancer, our understanding of development, dynamics and evolution of cells,

Clonal tracking is important because it gives a high-definition view of a population, and allows one to tease out the important parts of a complex group. To do this, sensitivity and complexity are two of the most important factors. A sensitive assay allows for the detection of clones that are present at a very low frequency, while a high complexity ensures that a large number of clones can be identified and tracked simultaneously. This is an important improvement over most current methods, because otherwise you are forced to either physically separate cells you are interested in, which perturbs the system you are studying, removes the cell from its microenvironment and from the signals and feedback afforded by neighbor cells, or you maintain the population dynamics and take ensemble measurements, which are averages of the members of the population, and may not accurately represent the status of any single cell, or represent the behavior of the predominant clone, while obscuring the contributions of less abundant clones.

Ideally, we would like to be able to trace the progeny of a single cell while not disturbing the population, to be able to take measurements and observe the behavior of a clone and to be able to do this tens or hundreds of thousands or even millions of times over within a single population, and to watch the dynamics of clonal relationships and the evolution of clones over time within the population, and in response to perturbations or challenges.

This information has both applications and implications for a number of fields and areas of study, including oncology, stem cell biology, and gene therapy. For example, tumors were once considered homogeneous clonal masses of cells, but are now increasingly recognized to be composed of a number of heterogeneous clones with different genetic mutations, growth rates, and tumor-seeding capabilities. Likewise, metastases may contain different clonal contributions, or even completely different clones, and therefore critical attributes, such as drug sensitivity, from the parent tumor. Heterogeneity among the cells of a primary tumor may be the source for the emergence of drug resistant clones leading to treatment failure or relapse. The contributions of clones to both the pre-and post-treatment tumor populations is vital information for our understanding of the phenomenon of therapy resistance and disease relapse, as these are often much more refractory to treatment and difficult to cure.

Stem cells are the vital members of a tissue, responsible for the maintenance of a tissue type, by asymmetric cell division to produce more restricted progenitors that in turn give rise to the mature, terminally differentiated cell types of a tissue. Embryonic stem cells give rise to all of the cell types of an organism, while more specialized stem cells are responsible for maintenance of a specific tissue, such as skin, muscle, or the hematopoietic system. These important cells are generally rare and physically dispersed amongst more restricted progenitors and terminally differentiated cell types, making them difficult to study. Being able to identify a true long-term repopulating stem cell requires the ability to track the progeny of a candidate cell to determine if it is truly multipotent and distinguish a stem cell from a more committed cell type. Currently, the standard for interrogation of a putative stem cell generally involves physical isolation of a single cell

either in a mouse or in an in vitro assay designed to demonstrate "stemness." Unfortunately, disturbing the cell's microenvironment and isolating it from other cells may have detrimental effects on the results of these assays.

Perhaps nothing has highlighted the necessity for effective clonal tracking methods like the events of the past decade in the gene therapy field. Gene therapy techniques seek to treat diseases at the genetic level, either by introducing a therapeutic transgene into targeted cells or by directly altering the sequence of a disease-causing allele (Naldini, 2011). Several human gene therapy trials have been completed with varying levels of success. However, the outgrowth of clones of treated cells has been seen on several occasions (Hacein-Bey-Abina, 2003; Schmidt, 2003; Mitsuhashi, 2007; Schwarzwaelder, 2007; Howe, 2008; Cartier, 2009; Hayakawa, 2009; Cavazzana-Calvo, 2010; Stein, 2010; Wang, 2010; Adair, 2012). This clonal growth resulted from mutations caused by the gene therapy treatment, which is most often a therapeutic transgene retrovirally introduced into the genomes of patient cells. These clones became dominant within the population due to growth advantages afforded by this mutagenesis. Three results were generally seen from these situations: (1) the dominant clone(s) died out, leading to treatment failure, or (2) a few times the dominant clones seemed to stabilize after a period of expansion, or (3) acquired additional mutational 'hits' which led to full oncogenic transformation, an unforeseen and unintended side effect of these early trials. Unfortunately, standard safety tests performed prior to these human trials were unable to detect the true risks of these methods. These events have demonstrated the necessity and utility of a sensitive and quantitative method for prospectively

measuring the effects and safety of any gene therapeutic on cells before its use in a patient.

REVIEW OF THE LITERATURE

The importance of detecting differences between two clones and determining the kinship of cells within a population has long been recognized and has inspired a number of methods and a large body of work over the past several decades. Often these methods have been limited by the scientific knowledge and technology available, and have evolved as technological capabilities have improved. General requirements of any clone tracking system include a method to stably and heritably mark the cell(s) of interest to ensure that this mark is passed down to all daughter cells, and a way to detect these marks for a readout of clonal behavior.

Tracking marker chromosomes

Some of the first cell tracking experiments were made possible by the identification of a strain of CBA mice that carried an extra small chromosome (Ford, 1956). Cells from these CBA/T6 mice transplanted into wildtype CBA mice were distinguishable from host cells by the presence of this marker chromosome visible in metaphase spreads. While not able to resolve the progeny of single cells, these experiments were able to show that the population of transplanted cells contained hematopoietic stem cells by their ability to reconstitute the blood systems of lethally irradiated recipient mice. Similarly, other experimental methods used to distinguish

transplanted cells from those of the host included transplants between male and female animals (Smith, 1991), and between mouse strains which expressed different hemoglobin alleles, identifiably by their migration patterns on an electrophoresis gel (Harrison, 1988). Radiation-induced chromosomal rearrangements provided another method for creating traceable cell clones in the laboratory, and were used to demonstrate the clonality of spleen foci in vitro (Abramson, 1977) and in vivo (Wu, 1968). However, the DNA damage used to mark these cells likely may have altered their behavior and potential, and the effectiveness of these techniques was limited by poor specificity and sensitivity.

Tracking viral integrations

With the advent of recombinant virus technology, key improvements in heritable cell marking became possible. Replication-incompetent recombinant viruses were engineered to deliver marker genes such as fluorescent proteins or antibiotic resistance for clonal selection and identification. The most common viruses for cell tracking studies are retroviruses, a class of RNA viruses leave a heritable mark on a host cell by integration of their genetic information into the host genome. Retroviral integration has been shown to be biased toward actively transcribed genes (Pryciak and Varmus, 1992; Neil and Cameron, 2002; Biasco, 2011), disruptions of which are more likely to changes in cell function. Because of this propensity, retroviral integrations are a common method of mutagenesis in genetic screens, and their potential effects on cells of interest should not be discounted. More recently, a class of retrovirus based on the Human Immunodeficiency Virus (HIV) has been developed for use because of its improved safety profile (Dull, 1998; Zufferey, 1998). Due to the semi-random nature of viral

integration, a transduced cell is uniquely marked by the location and pattern of integration sites(s). Over the years, numerous methods for detecting the unique junction of host and viral DNA at the integration site have been developed.

Southern blots have been used to track virally marked clones by detecting the sizes of clone-specific virus/host chimeric DNA fragments created by restriction enzyme digestion (Korczak, 1988; Capel, 1990; Jordan and Lemischka, 1990; Drize, 1996; Mazurier, 2004; McKenzie, 2006). However, the accuracy and sensitivity of Southern blots are limited by several considerations. A large amount of starting material is needed, making this method inappropriate for interrogating small samples. Integration sites too distant or near the chosen restriction sites will render the clone lost, and multiple bands of the same approximate size will not be resolved. Additionally, relatively rare clones are not detected, as 5-10% clonal contribution is the minimum for detection by Southern blot. This method is able to track a few clones and detect major clonal changes to a population, but are not able to detect all clones or more than a few clones at a time.

Multiple methods to improve the sensitivity of clonal detection of virally marked cells have been made over the years. Various PCR methods have been used to isolate viral integration sites for sequencing, including inverse PCR (Silver and Keerikatte, 1989; Nolta, 1996; Schmidt, 2001b), ligation-mediated (LM-PCR) and linear-amplification-mediated (LAM-PCR) (Schmidt, 2001a; Gentner, 2003; Kustikova, 2005; Harkey, 2007; Ciuffi, 2009; Gabriel, 2009; Stewart, 2010). These techniques generally require digestion of genomic DNA followed by fragment capture by ligation of adapter(s) and subsequent rounds of PCR amplification with primers to the adapter and to the known viral genome. As with other methods requiring digestion of genomic DNA, these

methods suffer from an inability to detect all integration sites due to restriction site proximity.

In all, these methods were improvements upon the poor sensitivity and large starting material requirements of Southern blot techniques, but were still plagued by inabilities to detect all clones in a population or precisely and accurately measure clonal contributions.

Tracking fluorescent markers

In addition to marking host genomic DNA, integrating viruses have been used to introduce selectable markers such as fluorescent proteins into cells. A few studies have tracked clones by infecting cells with a mix of viruses expressing different fluorescent proteins (Weber, 2011; Weissman, 2011). Cells expressing combinations of these proteins can be differentiated by the colors formed by these mixes. Up to 64 different clones can be identified in this way, making this technique best suited for identification of clones within static tissues. A non-viral method for identifying cellular clones uses fluorescent labels is FISH (Fluorescence in situ hybridization) (Anderson, 2011), which uses DNA probes conjugated to different color molecules to mark cells with specific genetic features, such as clone-specific chromosomal rearrangements, insertions or deletions. Unfortunately, this method cannot identify clones prospectively, and is limited by the number of features that can be detected in a single cell.

Tracking Single Cells

In order to avoid the potential mutational effects of viral integration and the complications that may pose for analysis of results, and with a lack of useful alternatives, studies were done on physically isolated single cells. The frequency of hematopoietic stem cells from populations sorted for various cell surface markers was determined by transplanting single cells into lethally irradiated mice and serial transplantation (Smith, 1991; Sieburg, 2006; Dykstra, 2007; Roeder, 2008; Sieburg, 2011). This method is both costly and time consuming, and required a huge number of mice. These methods were also hindered by the fact that the isolation and manipulation of the single cell may have altered its reproductive potential and "stemness."

Tracking genomic alterations

As genomic technologies improved, new methods for interrogating the genetic makeup of populations of cells were implemented. Copy number analysis (CNA) and array comparative genomic hybridization (CGH) have both been used to detect genetic changes to a population of cells, and can detect major clones with some success, but are very limited in their ability to detect clones that do not represent a significant proportion of the population (Mullighan, 2008; Mullighan, 2011; Zhang, 2012). Genomic differences within a population have more precisely detected by direct DNA sequencing, which can be targeted like with analysis of genomic hotspots or exome sequencing (Carlson, 2012), or more broad, with SNP (single nucleotide polymorphisms) detection and whole genome sequencing (Mullighan, 2011; Gerstung, 2012; Hou, 2012).

Tracking DNA barcodes

In the mid to late 1990s, a group published their method for tracking clones tagged with a unique library of retroviruses (Golden, 1995; Cepko, 1998; Satoh and Fekete, 2009). Their viral library, named CHAPOL, consisted of a library of avian vectors, each of which contained one of approximately 1000 different DNA sequences cloned from a pool of degenerate oligos. While this method greatly improved the number of clones which could theoretically be tracked, the system was not high throughput, as each cell had to be isolated, grown, and sequenced individually. Later methods included creating other randomized libraries of viruses, which were assayed in differing ways. Later studies presented detection of these sequences by microarray (Schepers, 2008; van Heijst, 2009), and while microarray-based detection of library members made this system more high-throughput, it was hampered by issues with sensitivity and was unable to accurately quantify the contributions of various clones.

In 2010, a group from the Netherlands published their study on the clonality of mouse HSC transplants using donor cells transduced with a retroviral library composed of a few hundred random "barcodes": short random sequences within the viral vector that distinguished one infected cell from another and were detectable by specific PCR amplification and traditional Sanger sequencing (Gerrits, 2010). The utility of the presented system was limited by the low complexity of the barcode library and small fraction of donor cells that were tagged, as well as the low-throughput of sequencing clones individually.

An improvement of this technology was published by Irving Weissman's group at Stanford University in 2011 (Lu, 2011). This group created lentiviral barcode libraries

with higher complexities, and coupled the readout with Illumina next-generation sequencing technology to greatly improve the sensitivity and number of clones that could be detected. Because of the nature of mouse transplant experiments, barcoded cells could not be enriched before introduction into recipients, so once again the majority of cells transplanted in this study were not marked, and each mouse received less than 2000 barcoded cells out of millions.

EXPERIMENTAL PLAN & HYPOTHESIS

We hypothesize that cell lines and other populations of cells are made up of heterogeneous clones that differ in genotype and phenotype and are dynamic both individually and in relationship to the population as a whole. In order to test this hypothesis, a minimally disruptive method for sensitively and quantitatively tracking cell clones within a large population is required. I have developed a novel clonal tracking system by coupling lentiviral delivery of a short DNA barcode to allow stable cell marking with a high-throughput deep sequencing method for detecting those barcodes. This system has the power to reveal the contributions of thousands of individual clones and track clonal dynamics and evolution within the entire population over time.

CHAPTER TWO
Creating a Complex Barcode Plasmid Library

BARCODE LIBRARY DESIGN & CONSTRUCTION

Introduction

In order to be able to stably, heritably, and detectably mark a large number of cells, we construct a pool of lentiviral vectors which each differ at just 20 base pairs, a DNA "barcode" that allow the identification of a marked cell by targeted sequencing of this short region. Utilizing Illumina deep sequencing technology, we are able to identify and quantify the frequencies of the entire population of barcodes in a high-throughput manner.

For the backbone of the barcode vector, we chose pLRG7, a third-generation, replication-incompetent, self-inactivating lentiviral expression vector from the lab of David Baltimore (Cal Tech) (Figure 2.1). This vector expresses GFP from a UBC promoter. The UBC promoter is ideal for this project because it is an endogenous promoter, which provides high levels of transgene expression in a wide variety of cell types, and which is less likely to be silenced by the viral DNA silencing pathways in transduced cells.

Within this vector, we identified a region upstream of the GFP expression cassette (to avoid expression of the non-coding barcode sequence) with unique restriction sites (XhoI and BamHI) for cloning of the barcode library. Into this site, we cloned the full Illumina P5 genomic sequencing adapter sequence:

AATGATACGGCGACCACCGAGATC

TACACTCTTTCCCTACACGACGCTCTTCCGATCT, followed by an additional unique restriction site (AscI) to use for insertion of the barcode fragment. The inclusion of the P5 sequence serves three purposes: (1) It allows us to use a single step PCR to amplify the barcodes from the template DNA and prepare them for Illumina sequencing, reducing the amount of skewing to the library from handling and multiple manipulations, including adapter ligation. (2) Inserting the P5 sequence at this position allows us to maximize the small number of bases obtained from each Illumina read, and (3) allows us to sequence only the barcode end of our PCR fragments by always having the barcode immediately next the first sequencing primer used, which is not a possibility when using the traditional Illumina sequencing preparation method of ligating Y adapters onto the target DNA.

We chose to make the barcode sequence 20 bases long in order to best utilize short 36bp single-end Illumina reads while ensuring a great excess of the number of possible sequences to prevent having two identical or very similar barcodes occur independently in the same library. The number of possible barcode sequences is calculated as 4 to the 20th power, or approximately 1.1 trillion.

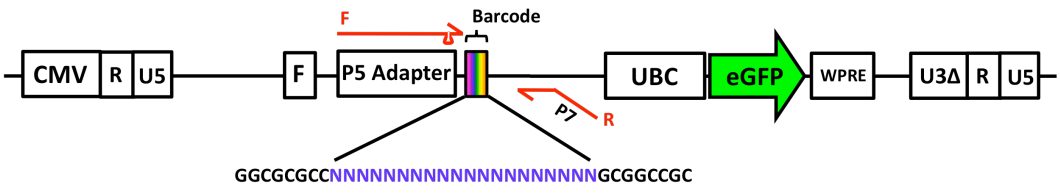


Figure 2.1. Barcode Lentiviral Vector.

The pLGR7-based lentiviral barcode vector features a hybrid CMV-R-U5 5' LTR, a Flap element (F) upstream of the P5 Illumina adapter and 20 base pair barcode sequence. The construct contains an Ubiquitin C Promoter driving an eGFP gene. Expression of eGFP is enhanced by a downstream woodchuck promoter response element (WPRE) flanked by a delta U3 self-inactivating 3' LTR. Forward (F) and reverse (R) primer binding positions are indicated, and the 4 bp bubble created by the multiplexing tag is shown.

Materials and Methods

To create the library of barcodes to use for cloning, we synthesized two oligonucleotides:

Oligo #1: 5'- AGT CGG CGC GCC NNN NNN NNN NNN NNN NNN NNG CGG CCG CCC TGC AGG GGA TCC AGT C -3'

Oligo #2: 5'- GAC TGG ATC CCC TGC AGG GCG GCG CC -3'

The key parts of these oligos are described below:

Oligo #1:

1. 4 randomly chosen bases to help aid in complete digestion by restriction enzymes, which are often less effective when their recognition sites are positioned at the very ends of DNA.
2. 5' restriction enzyme recognition site – AscI: GGCGCGCC (for cloning)
3. 20 'N's, allowing any of the four DNA bases to be synthesized at these positions. This makes up the random barcode sequence.

4. 3’ restriction sites – NotI: GCGGCCGC SbfI: CCTGCAGG BamHI: GGATCC
5. 4 more randomly chosen bases

Oligo #2: 26 bases complimentary to the 3’ end of Oligo #1 to create a double strand barcode fragment.

These oligos were annealed by mixing equimolar amounts in the presence of 1X Phusion High Fidelity Buffer (New England Biolabs) and heating in a 95° C heat block for five minutes, then allowing the entire block to cool to room temperature slowly. Annealed oligos were extended by mixing with 2X Phusion Polymerase High Fidelity Master Mix (New England Biolabs) and incubating at 72° C for 30 minutes. Phusion Polymerase was chosen because of its very low error rate during DNA synthesis, and master mix was used to minimize pipetting steps and sources of contamination. Double stranded barcode fragments were prepared for cloning by digestion with AscI and BamHI and ligated into the similarly digested vector at a molar ratio of 10:1.

Ligations were pooled, ethanol precipitated using Pellet Paint (EMD Millipore), and electroporated into ElectroMAX DH10B electrocompetent *E. coli* (Invitrogen). After recovery for one hour in SOC, 5% of the reaction was spread on an ampicillin selection plate and incubated overnight at 37° C, while the remainder of the cells were inoculated into 200mL rich bacterial growth medium (SOC) with 100ug/mL ampicillin and grown overnight at 37° C with shaking. Plasmid DNA was isolated from the cells grown in liquid culture after 16 hours with an Endotoxin-free Maxi Prep Kit (QIAGEN). 16 individual colonies were picked from the ampicillin plate and sequenced by Sanger sequencing.

Results

Complexity of the plasmid barcode library was calculated by multiplying the number of colonies on the plate (1066) by the dilution factor (1:19) for an estimated complexity of 20,254. Of 16 individually sequenced colonies, 15 contained a single barcode, while 1 contained a concatemer of 3 barcodes. Each of the sequenced barcodes was unique within this small cohort.

Discussion

These results demonstrate the importance of using the lowest ratio of insert to vector that gives sufficient clones in order to minimize ligation of multiple barcode fragments together into a single vector. With the methods described above, we were able to create a library of unique, random barcodes with an estimated complexity of 2×10^4. These methods are easily adjustable to create barcode libraries of almost any complexity and in any type of vector.

BARCODE LIBRARY DEEP-SEQUENCING AND ANALYSIS

Introduction

In order to measure the actual complexity of the plasmid barcode library and to determine the reproducibility of the barcode system and to detect what effect sample handling, preparation, and sequencing have on the observed frequencies of barcodes, four independent barcode amplification PCR reactions of the plasmid barcode library were prepared and sequenced. The tens of millions of lines of data generated by a single lane of Illumina sequencing necessitate specialized bioinformatics methods for data handling and analysis. We have been fortunate to collaborate with David Mittelman's group at Virginia Tech University, who have written the programs necessary to effectively analyze the data created by this barcode system.

Materials and Methods

PCR amplification and sequencing preparation of barcodes

Phusion master mix	25uL
Forward Primer (30uM)	1uL
Reverse Primer (30uM)	1uL
DMSO	1uL
Template DNA	XuL
Sterile water	to 50uL

Hot start PCR
Initial denaturation step: 98°C – 2 minutes

26 Cycles: 98°C – 10 sec.
62°C – 15 sec.
72°C – 20 sec.
Final Extension: 72°C – 5 minutes

Forward primer sequence:

5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC TNN NNG GCG CGC C -3'

Reverse primer sequence:

5'- CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC TCG CTA TGT GTT CTG GGA AAT CAC C -3'

The PCR(s) for each sample were run on separate 1.1% TBE gels at 100V, and the expected 250bp fragment was excised from the gel and purified with the QIAquick Gel Extraction Kit (QIAGEN). Size and concentration of each sample was confirmed on a 2100 BioAnalyzer (Agilent). Sample concentrations were adjusted to 10uM in TE buffer, then pooled and submitted for single-end, 36-cycle deep sequencing on an Illumina Genome Analyzer II by the Stanford Functional Genomics Core Facility (sfgf.stanford.edu).

Analysis of Sequencing Data

Results from the Illumina sequencing runs were provided in fastq format, which provides information about the confidence in the accuracy of each base call, called a Phred score or quality score. Quality scores range from 0 to 35. Scores above 28 are

generally considered very good quality and those above 20 are of reasonable quality; below 20 the accuracy of the base call drops significantly.

The steps taken by the barcode counting and clustering program written by David Mittelman and Gareth Highnam (Virginia Tech) are outlined in Figure 2.2. The input fastq file is first sorted into component samples by 4bp multiplexing tags at the beginning of each sequence. Sequences without the correct sequences before and after the barcode are then removed, and any base within the 20bp barcode with a quality score below the cutoff value is replaced with an 'N'. Any sequence containing more than the indicated number of Ns is discarded from further analysis. Sequence number for each barcode is counted, and barcodes that differ by less than a set number of bases (mismatches) are considered to be the same and are clustered together. The resulting barcode cluster is named based on the consensus of the sequences contributing to that cluster. And additional program allows comparison of the frequencies of each barcode between two or more samples, grouping barcodes with three or fewer mismatches together. This program will be available for download from <https://github.com/adaptivegenome/clusterseq> upon publication.

The resulting comma delimited file contains the count for each barcode in each of the input samples. Frequency for each barcode in each sample is calculated as percent of the population. Sequence background noise is greatly reduced by the filtering steps and quality checks made during the program analysis. Final background noise removal steps include replacing any barcode frequency below the possible limits based on using 300,000 cells' worth of genomic DNA for barcode amplification for sequencing prep, or

less than 0.0003%, with a zero, and removing any barcodes that do not meet this minimum frequency in at least 10% of the samples being compared.

The quality score cutoff and numbers of allowed Ns and mismatches was determined by testing a matrix of combinations on a single data set to determine the conditions that seemed to give the best results. It is important that the program parameters be such that barcodes differing at just a few bases due to PCR or sequencing errors are recognized and clustered together, while not being so permissive as to cluster unrelated barcodes or noise.

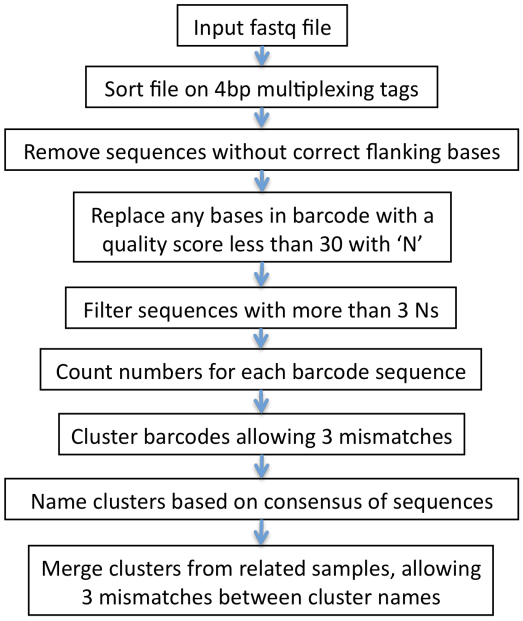


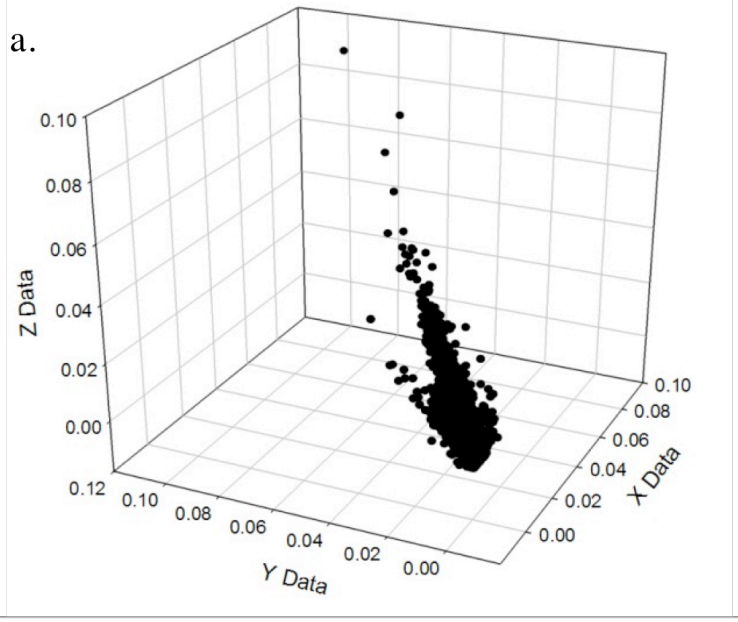
Figure 2.2. Workflow of Barcode Counting Program.

Results

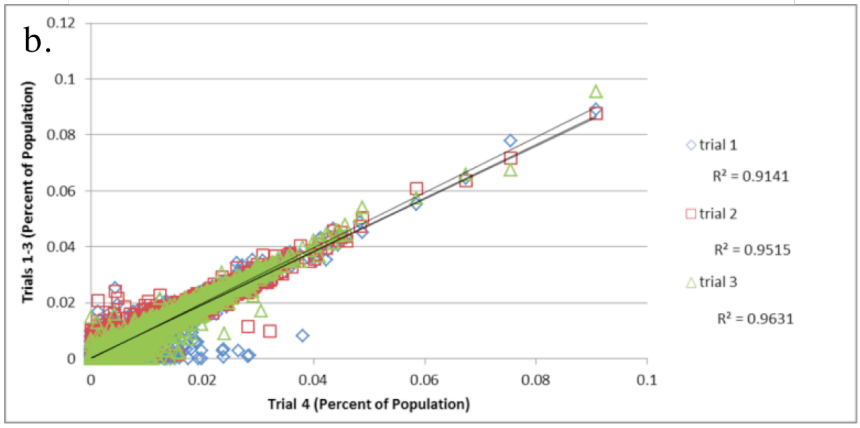
Four independent PCR barcode amplification and sequencing preparations of the barcode plasmid library were made and submitted for deep sequencing. Each of the resulting sequencing files was analyzed with the following parameters: a minimum quality score of 30, up to 3 Ns allowed per sequence, and up to 3 mismatches allowed between two sequences for clustering. The number of barcodes identified in the four sequencing replicates of the plasmid library ranged from 12,482 to 12,887, and averaged 12,674, with a standard deviation of 175.

In order to compare the similarity of the replicates at each barcode, frequency across three replicates were plotted on 3D and on 2D scatter plots (Figure 2.3). The r-squared values between any two replicates were each greater than 0.9, indicating a high degree of similarity between the replicates. As shown in Table 2.1, 12,173 barcodes were present in all 4 replicates, 1122 were shared among 3 replicates, 330 were found in half of the samples, and a combined 220 barcodes were present only in a single sample. The library was over-sequenced a minimum of 460 times in each of the four samples. The average frequency of a barcode in the library is approximately 0.008%, or 1 in 12,500, demonstrating the even distribution of barcodes. The sequencing results for the barcode plasmid library also revealed a bias in the base composition of the randomized 20bp barcode sequence. The average base composition across the barcode sequence is 42.15% C, 21.95% A, 18.35% G, and 17.55% T. This bias is known to occur when randomized bases are machine mixed during oligo synthesis, and can be avoided by hand mixing the bases to equal proportions.

a.


Z Data
0.10
0.08
0.06
0.04
0.02
0.00
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Y Data
0.10
0.08
0.06
0.04
0.02
0.00
X Data

b.


Trials 1-3 (Percent of Population)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.02
0.04
0.06
0.08
0.1
Trial 4 (Percent of Population)
trial 1
R² = 0.9141
trial 2
R² = 0.9515
trial 3
R² = 0.9631

c.

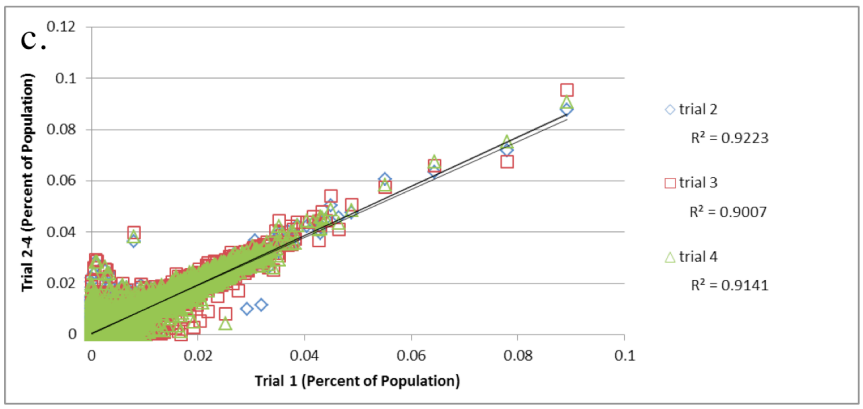

Trial 2-4 (Percent of Population)
0.12
0.1
0.08
0.06
0.04
0.02
0
0
0.02
0.04
0.06
0.08
0.1
Trial 1 (Percent of Population)
trial 2
R² = 0.9223
trial 3
R² = 0.9007
trial 4
R² = 0.9141

Figure 2.3. Comparison of Barcode Plasmid Library Sequencing Replicates.

(a.) Trials 1 (X-axis), 2 (Y-axis), and 3 (Z-axis) were plotted in Sigma Plot. (b.) Trials 1-3 were plotted against Trial 4, and linear regression and r-squared values were calculated for each pair-wise comparison. (c.) Trials 2-4 were plotted against Trial 1, and linear regression and r-squared values were calculated for each pair-wise comparison.

	Total Barcodes	Unique (1/4)	Common (4/4)	3/4	2/4
Trial 1	12730	49	12173	1122	330
Trial 2	12887	135			
Trial 3	12482	21			
Trial 4	12595	15			
Total	13098				
Average	12673.5				
Std. Deviation	174.7				

Table 2.1. Shared and unique barcodes across four sequencing replicates of the plasmid library.

The total number of barcodes found in each sample after the removal of background noise, the numbers of barcodes found only in a single sample, any barcodes found in any two, three, or all four of the samples.

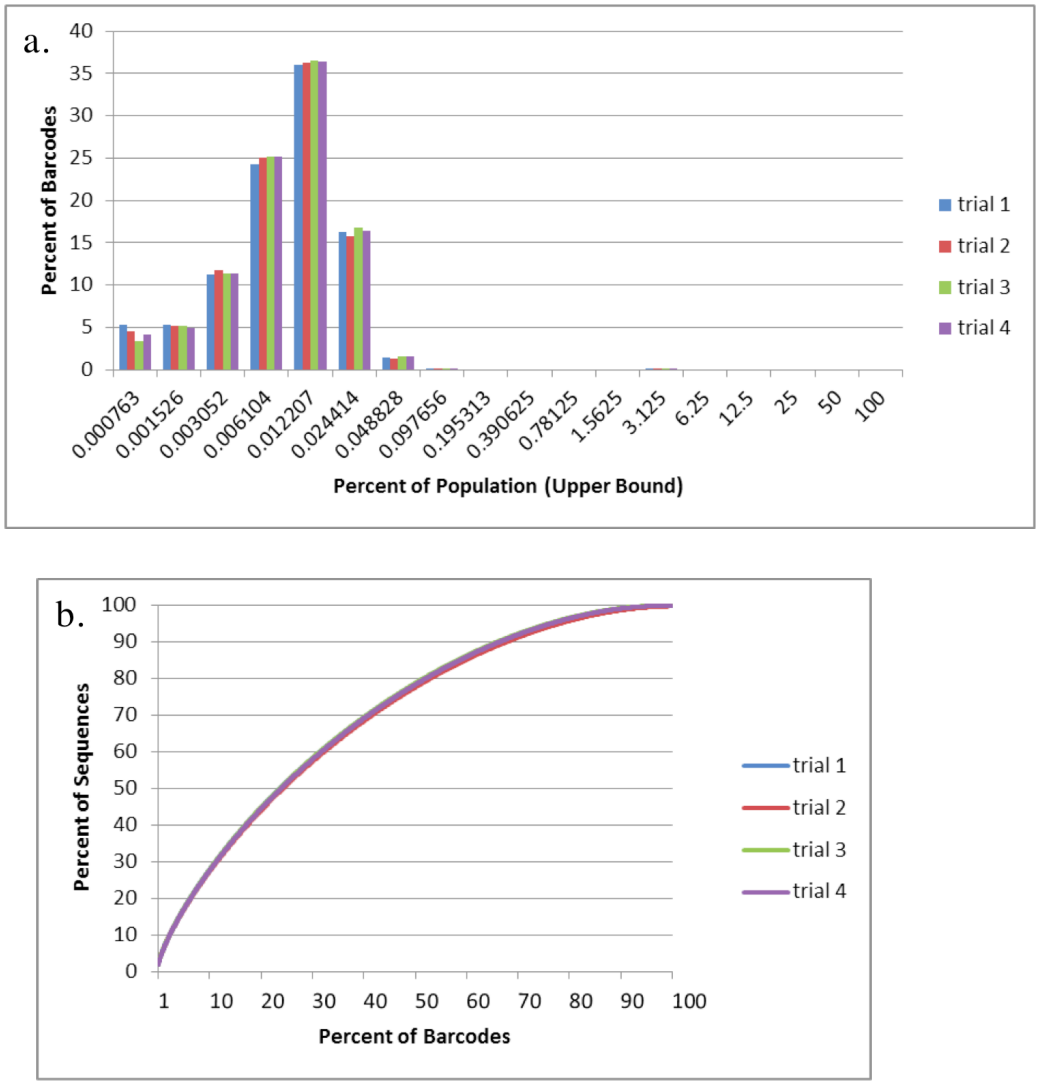


Figure 2.4. Distribution of the barcode plasmid library.

(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency).

	% Sequences by Barcodes			% Frequency (Count/10^5)	
	top 25%	**top 50%**	**top 75%**	**Median**	**Average**
Trial 1	51.3	77.9	93.9	0.006762 (6.8)	0.008041 (8.0)
Trial 2	50.9	77.5	93.6	0.006538 (6.5)	0.007731 (7.7)
Trial 3	50.8	77.3	93.4	0.006701 (6.7)	0.007993 (8.0)
Trial 4	50.7	77.3	93.5	0.006638 (6.6)	0.007921 (7.9)
Average	50.9	77.5	93.6	0.006659 (6.7)	0.007922 (7.9)
Std. Deviation	0.228	0.245	0.187	0.000083 (0.1)	0.000118 (0.1)

Table 2.2. Plasmid barcode library statistics

The percent of the sequences from each sample that made up the 25, 50, or 75 percent most frequent barcodes. The more of the sequences contained by the top barcodes, the less evenly distributed the population. Also listed are the median and average frequency (percent of population represented by each barcode), for the population at each timepoint. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.

Discussion

The results from the sequencing of the barcode plasmid library are critical for validating the system and for demonstrating the reproducibility of sample preparation and sequencing. The actual complexity of the library was determined by oversequencing to be 12674. This is lower than the complexity of 20,000 calculated during library cloning. This may be due to the inflation of the original calculated complexity from non-barcoded vector and other cloning artifacts or from differences in survival of *E. coli* on a plate compared to in liquid culture.

With these results we demonstrate that we are able to create complex, fairly evenly distributed libraries of plasmids containing random 20 base pair barcodes and that our methods for sample preparation and the Illumina sequencing process itself do not introduce significant alterations to the observed frequency of individual barcodes within the library. Several features of our system may have helped to avoid the bias and skewing observed with other uses of PCR amplification or Illumina sequencing technology. Because all of the barcode fragments are the same size, and have the exact same base composition at 230 of 250 bases, our system avoids biases due to amplicon size during PCR, as well as the effects of large differences in GC content.

By over-sequencing the barcode library by a minimum 460 fold per replicate, we were able to gain confidence that we had an accurate measure of the true complexity of the barcode library. The measured complexity of 12,674 is , and will allow us to track many more clones with a level of precision and sensitivity unique to this system.

We initially chose to multiplex our samples as a cost effective measure to use fewer lanes of sequencing and because a single sample did not require 30-40 million

reads to be sufficiently over sequenced. However, our multiplexing tags became a critical part of the experimental design, due to the design of our barcode fragments and the base-reading software used by the Illumina sequencing technology. Because all of our barcode fragments have the same sequence before and after the barcode region, and because these fixed bases are immediately after the integrated P5 Illumina adapter sequence used for priming during sequencing, these bases would be the first sequenced if not for the multiplexing tags. Having a sea of a single base type on the flow cell during each of the first few sequencing cycles causes problems for Illumina base recognition software, leading to abysmal read quality. By using 4bp multiplexing tags with an even distribution of each base at each of the four positions, and pooling four complimentary multiplexing tags per lane of sequencing, the software is able to accurately identify each of the four bases in each of the first four cycles, preventing problems with the homogeneous bases during later cycles.

CHAPTER THREE
Barcodes in Common Cell Lines

INTRODUCTION

In order to validate that the barcoding system for use in tracking cellular clones, we used barcode lentivirus to create cellular barcode libraries (Figure 3.1) in three widely used cell lines: HeLa, HEK-293T, and K562. In order to get as an accurate idea of the clonal dynamics of these cells, it was important to avoid any introducing any artificial sources of skew or bias. This was accomplished by maintaining culture conditions as close to ideal as possible. Cells were kept in log-phase growth and never allowed to become confluent. 3 x 10^5 cells were passaged at each split in order to avoid an artificial bottleneck of the population. Cellular barcode libraries were passaged in triplicate (Figure 3.2) in order to determine whether clonal dynamics were set in the original population as would be indicated by similarity of the clonal make up of the populations after 90 population doublings, or if clonal evolution occurs during relatively short passages in vitro, leading to divergent populations with different major clones.

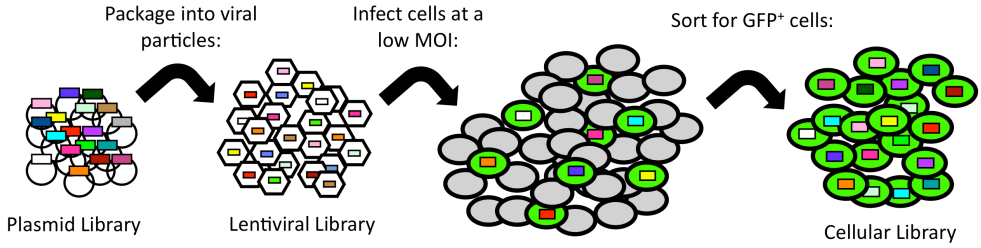


Figure 3.1. Creating Cellular Barcode Libraries.

Schematic representation of the steps involved in creating a cellular barcode library from the plasmid barcode library. The plasmid library is packaged into lentivirus, which is

then used to infect cells at a low MOI to prevent multiple infections to a single cell. Transduced cells expressing GFP are then sorted to form the cellular barcode library.

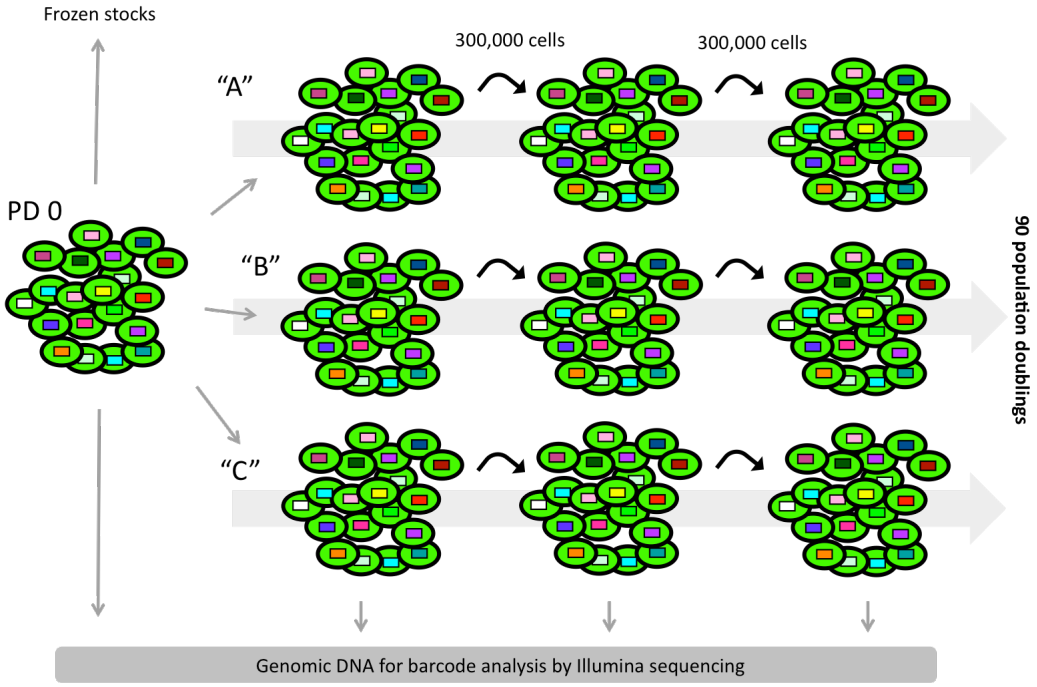


Figure 3.2. Diagram of in vitro cell passaging experiments.

Population doubling zero "PD 0" cells are divided into 300,000 cell populations, three of which are grown separately in parallel (A-C). Cells are passaged every 3 days, and 300,000 cells are subcultured at each passage, an approximately 1:10 split. Cells are harvested every 10 or 30 population doublings for barcode analysis by targeted deep sequencing (Illumina).

BARCODE LENTIVIRUS PRODUCTION

Introduction

In order to introduce the barcode library into cells, we used the barcode plasmid library to produce high-titer lentivirus. We used caffeinated media on virus producer cells, as our lab has shown that this increases lentiviral titer (Ellis, 2012).

Materials and Methods

On the day prior to transfection, HEK-293T cells were plated on gelatin-coated 10cm cell culture dishes at a confluency of approximately 40% in DMEM (Cellgro) supplemented with 10% bovine growth serum(Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. Just prior to transfection, media was removed and replaced with fresh media.

Transfection mix was prepared per 10cm plate for 18 plates as follows:

1. Mix 3ug pVSVG + 5ug pRRE + 2.5ug pRSV REV + 10ug barcode library plasmid in 2.5mM Hepes to a final volume of 250uL, then add 250uL 0.5M $CaCl_2$.
2. DNA mix was added to 500uL 2X HeBS, 1-2 drops per second while vortexing vigorously.
3. After letting stand undisturbed for 30 minutes, the mix was added (1mL/plate) drop wise slowly onto cells. Plates were gently tilted to distribute the transfection mix and were kept in a humidified, 37° C, 5% CO_2 incubator overnight.

4. 18 hours later, the transfection mix and media was removed from cells and replaced with 10mL fresh DMEM supplemented with 2% bovine growth serum, 100 units/mL penicillin, 100ug/mL streptomycin, 2mM L-glutamine, and 4nM caffeine.
5. 24 hours later, virus-containing supernatant was removed from the cells and centrifuged at 3000 RPM, 4° C for 15 minutes to remove cellular debris, then passed through a 0.45 micron filter and stored at 4° C.
6. Virus particles were concentrated by ultracentrifugation in a SW-28 swing bucket rotor at 25,000 RPM, 4° C, for 90 minutes. Supernatant was drained gently from the viral pellet, and viral particles were resuspended in 100uL PBS overnight at 4° C. Viral particles were fully resuspended by pipetting 10 times and 50uL aliquots were immediately stored at -80° C until use. Everything that came into contact with virus was treated with bleach prior to disposal.
7. Viral titer was determined by infecting 10^5 HEK-293T cells with a series of dilutions of the virus stock. Percent infected cells were measured after 48 hours by flow cytometry, and wells with less than 10% infection rates were used for titer calculations: (0^5 cells) x (%GFP+) x (uL of virus stock used for infection) = viral titer in infectious units per uL.

Results

Titer of the barcode lentivirus was performed as described above, and the number of GFP positive cells, measured 48 hours post transduction with 0.01uL virus, was an average of 9.1% across three replicates; therefore the viral titer was determined to be approximately 9 x 10^5 infectious units per microliter. Sufficient stock of this viral prep was stored for the experiments described hereafter.

BARCODE TRACKING OF HELA CELLS

Introduction

HeLa cells were chosen for analysis because of their status as the oldest and most commonly used human cancer cell line. HeLa cells were derived from the cervical cancer cells of a patient named Henrietta Lacks in 1951 and are arguably the most famous and widely-published cell line in the world (Scherer, 1953). It would be expected that after so many years in culture, the HeLa cell line will be homogeneous.

Materials and Methods

Creation of HeLa Barcode Library

HeLa cells were a kind gift of Dr. Alejandro Sweet-Cordero (Stanford). Log-phase cells were trypsinized, resuspended in a small volume of media (>10^6 cells per mL), and 2 x 10^6 cells were infected with barcode lentivirus calculated to achieve an MOI of 0.05. Cells were diluted in additional media and plated 1 hour after the start of infection. Transduction levels were confirmed by flow cytometry after 48 hours. 4 days after

infection, 2 x 10^5 GFP positive cells were isolated by flow sorting on a FACS Aria II (Becton Dickinson). This sorted population of barcoded cells were then expanded in culture for 5 days, then divided for various uses as population doubling zero (PD0) cells.

In vitro culture experiments

HeLa cells were maintained in DMEM(Cellgro) supplemented with 10% bovine growth serum(Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. 3 aliquots of 300,000 PD0 cells each were separated into individual 6-well tissue culture plates as populations A, B, and C. Every three days, cells were trypsinized, counted, and analyzed for GFP expression by flow cytometry. 300,000 cells from each population were transferred to a new well of a 6-well plate with 3mL of fresh media.

Sequencing sample preparation

Genomic DNA was harvested from pelleted cells with the DNeasy Blood and Tissue Kit (QIAGEN). PCR amplification of barcodes was performed as for the plasmid barcode sequencing in Chapter 2, except that the genomic DNA from 300,000 cells (approximately 5.6ug due to the aneuploidy of HeLa cells) was used as template starting material, divided between 8 50uL PCR reactions, which were subsequently pooled.

Results

Barcoded HeLa cells grew rapidly, the population doubling a little more than once every 24 hours, and the doubling rates of the three populations were very similar. (Figure 3.3). A small but steady decrease in the percent of GFP positive cells was

observed in all three populations during the last 30 population doublings of the experiment. This observance has not yet been explained, but may be the result of silencing of the GFP transgene in a few cells, or may indicate a growth advantage in one of the small percentage of non-tagged cells.

HeLa sequencing results were processed in the same as for the barcode plasmid library (Chapter 2). The number of barcodes in each population over time (Figure 3.4c) demonstrates the steady loss of clones during the progression of the experiment. In order to visualize the distribution of barcode frequencies within each timepoint, the frequency of each barcode was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 3.4a) shows that the distribution of barcode frequencies within the PD0 starting population was relatively normal, indicated by the bell-shaped curve, but over time most barcodes became less frequent as a few became overrepresented, as seen by the shift of barcodes in the PD90 sample to the lower frequency bins, while a few barcodes begin to appear in the higher frequency bins.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 3.4b). Again, this plot demonstrates that the barcodes in the earliest timepoint (PD0) are relatively evenly distributed (a perfect distribution would follow the line (y=x), but over the course of the experiment, a small fraction of the barcodes become more dominant, and are represented in a larger percent of the sequences from that sample. The median and average percent frequency of

barcodes at each timepoint was calculated (Table 3.1), and agrees with the other data showing a progression from evenly distributed to a separation between low-frequency clones and a few more dominant ones. The median barcode frequency decreases with time, while the average barcode frequency increases.

In order to look at behavior of individual clones, barcode frequency at each timepoint of the experiment was plotted (Figure 3.5). A sampling of every 300th barcode in the population (Figure 3.5a) gives an unbiased cross-section view of the clonal dynamics. The top twenty most frequent barcodes at each timepoint are plotted individually (Figure 3.5d-g) in order to show trends among the dominant clones. A selection of barcodes which disappeared from the population by PD60 or PD90 were plotted (Figure 3.5b) to determine whether these barcodes were simply the least frequent in the population and were lost due to stochastic reasons, or if they appeared to be actively selected against. Based on this data, it appears that the majority of clones were lost due to stochastic mechanisms and clonal competition, although other reasons for the disappearance of clones cannot be ruled out.

To determine whether the same clone behaved similarly in parallel populations, three randomly selected barcodes with different starting frequencies were tracked across each of the three populations (Figure 3.5c). These plots show that the divergence between populations increases over time, as the frequencies of each barcode at PD30 are much more similar than at PD90.

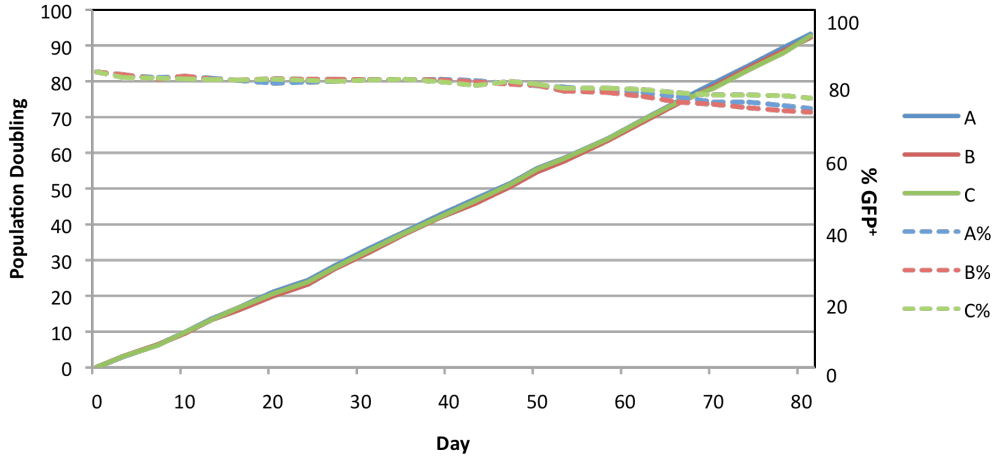


Figure 3.3. Growth and GFP levels of HeLa barcode library.

Population doublings of the HeLa barcode cell populations (solid lines) and the percent of the population expressing GFP (dashed lines), for the duration of the experiment.

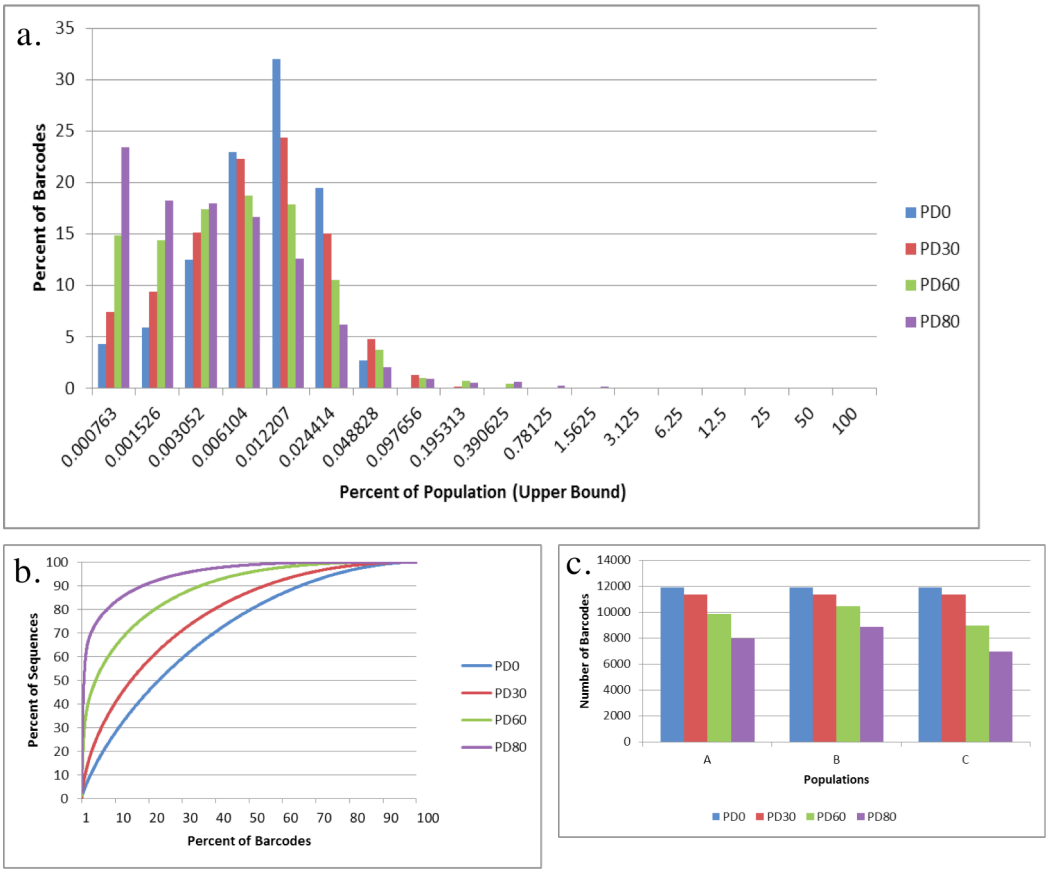


Figure 3.4. Complexity and distribution of clones in HeLa barcode experiments. (a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency). (c.) Number of barcodes in each population at each time point.

	Frequency (Count/3x10^5)	
Pop. Doublings	**Median**	**Average**
0	0.006749 (20.3)	0.008375 (25.1)
30	0.005391 (16.2)	0.008795 (26.4)
60	0.003446 (10.3)	0.010138 (30.4)
90	0.002118 (6.4)	0.012538 (37.6)

Table 3.1. Median and average barcode frequency over time in HeLa cells.

The median and average frequency (percent of population represented by each barcode), for the population at each timepoint were calculated. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.

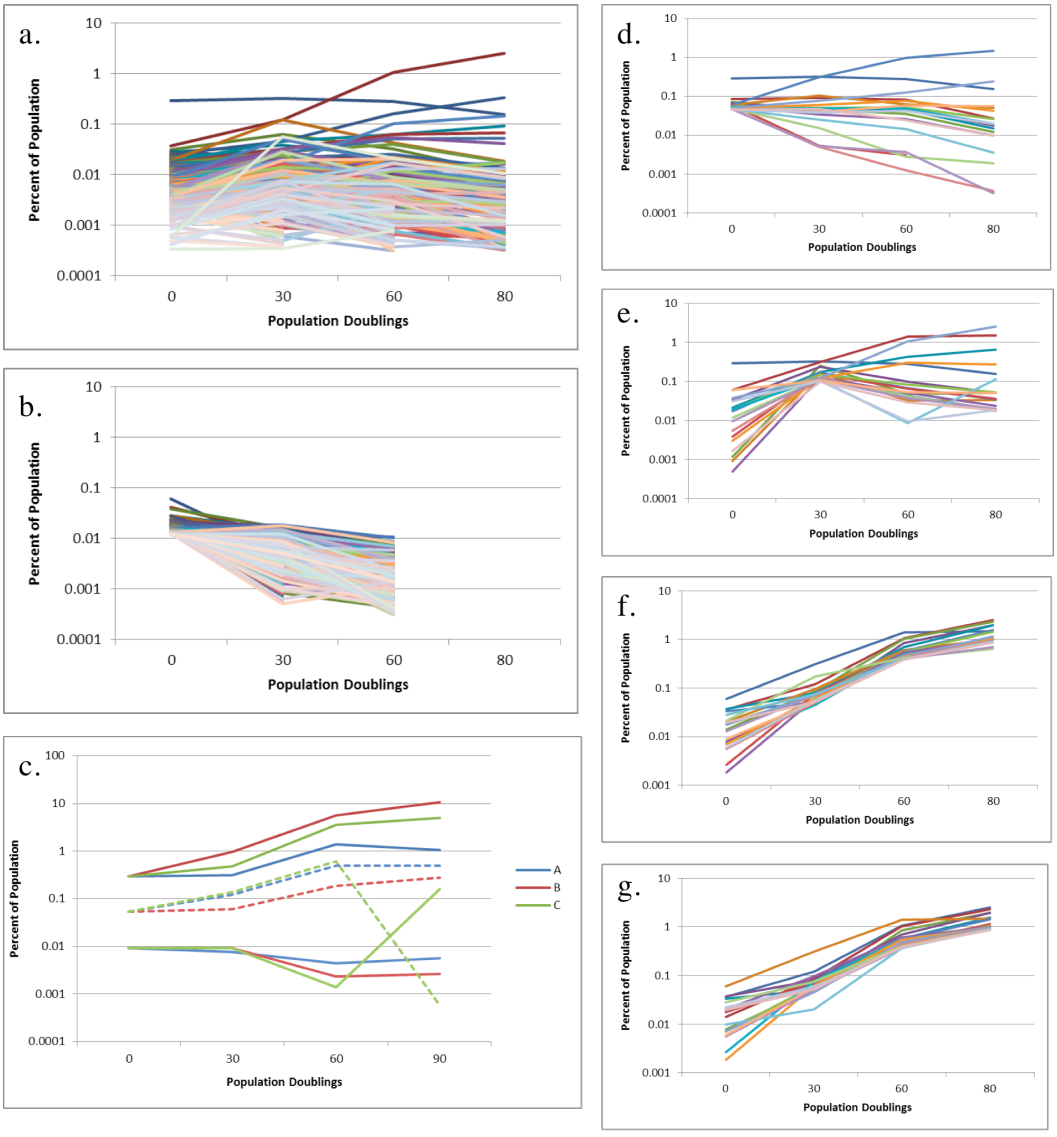

a.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
80
Population Doublings
b.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
80
Population Doublings
c.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
Population Doublings
A
B
C
d.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
80
Population Doublings
e.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
80
Population Doublings
f.
Percent of Population
10
1
0.1
0.01
0.001
0
30
60
80
Population Doublings
g.
Percent of Population
10
1
0.1
0.01
0.001
0
30
60
80
Population Doublings

Figure 3.5. Clonal Dynamics in HeLa cells.

Each line represents the frequency of a single barcode-marked clone over the course of the experiment. (a) a sampling of the behavior of 250 clones from a cross-section of the population, shown on a log scale. (b) barcodes that disappeared from the population after 30 or 60 population doublings. (c) Three randomly selected barcodes are tracked across all three populations (A, blue; B, red; C, green). (d) The top 20 most frequent barcodes at PD0. (e) The top 20 most frequent barcodes at PD30. (f) The top 20 most frequent barcodes at PD60. (g) The top 20 most frequent barcodes at PD90.

BARCODE TRACKING OF HEK-293T CELLS

Introduction

HEK-293T cells are an immortalized line of human embryonic kidney cells. These cells are widely used due to their ease of transfection.

Materials and Methods

Creation of HEK-293T Barcode Library

HEK-293T cells were obtained from ATCC (Logan, UT). Log-phase cells were trypsinized, resuspended in a small volume of media (>10^6 cells per mL), and 2 x 10^6 cells were infected with barcode lentivirus calculated to achieve an MOI of 0.05. Cells were diluted in additional media and plated 1 hour after the start of infection. Transduction levels were confirmed by flow cytometry after 48 hours. 4 days after infection, 2 x 10^5 GFP positive cells were isolated by flow sorting on a FACS Aria II (Becton Dickinson). This sorted population of barcoded cells were then expanded in culture for 5 days, then divided for various uses as population doubling zero (PD0) cells.

In vitro culture experiments

HEK-293T cells were maintained in DMEM (Cellgro) supplemented with 10% bovine growth serum(Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. 3 aliquots of 300,000 PD0 cells each were separated into individual 6-well tissue culture plates as populations A, B, and C. Every three days, cells were trypsinized, counted, and analyzed for GFP expression by flow cytometry. 300,000 cells from each population were transferred to a new well with 3mL of fresh media.

Sequencing sample preparation

Genomic DNA was harvested from pelleted cells with the DNeasy Blood and Tissue Kit (QIAGEN). PCR amplification of barcodes was performed as for the plasmid barcode sequencing in Chapter 2, except that the genomic DNA from 300,000 cells (approximately 5.0ug due to the aneuploidy of HEK-293T cells, Appendix A) was used as template starting material, divided between 8 50uL PCR reactions, which were subsequently pooled.

Results

Barcoded HEK-293T cells grew rapidly, the population doubling a approximately every 21 hours, and the doubling rates of the three populations were very similar. (Figure 3.6). HEK-293T sequencing results were processed as described previously (Chapter 2). The number of barcodes in each population over time (Figure 3.7c) demonstrates the generally small loss of clones during the progression of the experiment. In order to visualize the distribution of barcode frequencies within each timepoint, the frequency of each barcode was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 3.7a) shows that the distribution of barcode frequencies within the PD0 starting population was relatively normal, indicated by the bell-shaped curve, but over time most barcodes became less frequent as a few became overrepresented, as seen by the shift of

barcodes in the PD90 sample to the lower frequency bins, while a few barcodes begin to appear in the higher frequency bins.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 3.7b). Again, this plot demonstrates that the barcodes in the earliest timepoint (PD0) are relatively evenly distributed (a perfect distribution would follow the line (y=x), but over the course of the experiment, a small fraction of the barcodes become more dominant, and are represented in a larger percent of the sequences from that sample. The median and average percent frequency of barcodes at each timepoint was calculated (Table 3.2), and agrees with the other data showing a progression from evenly distributed to a separation between low-frequency clones and a few more dominant ones. The median barcode frequency decreases with time, while the average barcode frequency increases.

In order to look at behavior of individual clones, barcode frequency at each timepoint of the experiment was plotted (Figure 3.8). A sampling of every 300th barcode in the population (Figure 3.8a) gives an unbiased cross-section view of the clonal dynamics. The top twenty most frequent barcodes at each timepoint are plotted individually (Figure 3.8d-g) in order to show trends among the dominant clones. A selection of barcodes which disappeared from the population by PD60 or PD90 were plotted (Figure 3.8b).

To determine whether the same clone behaved similarly in parallel populations, four randomly selected barcodes with different starting frequencies were tracked across

each of the three populations (Figure 3.8c). These plots show that the clones generally behaved very similarly in the parallel populations.

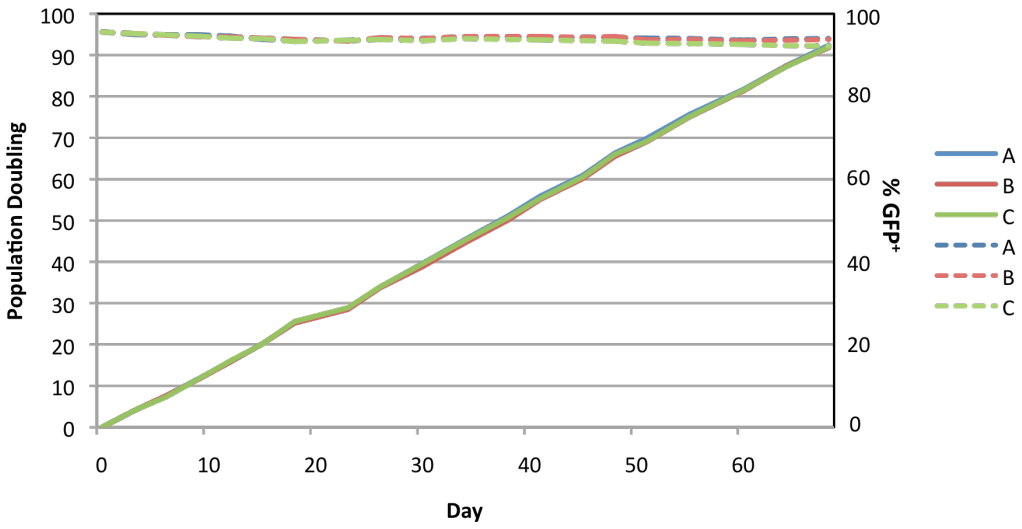


Figure 3.6. Growth and GFP levels of HEK-293T barcode library.

Population doublings of the HeK-293T barcode cells (solid lines) and the percent of the population expressing GFP (dashed lines), over the duration of the experiment.

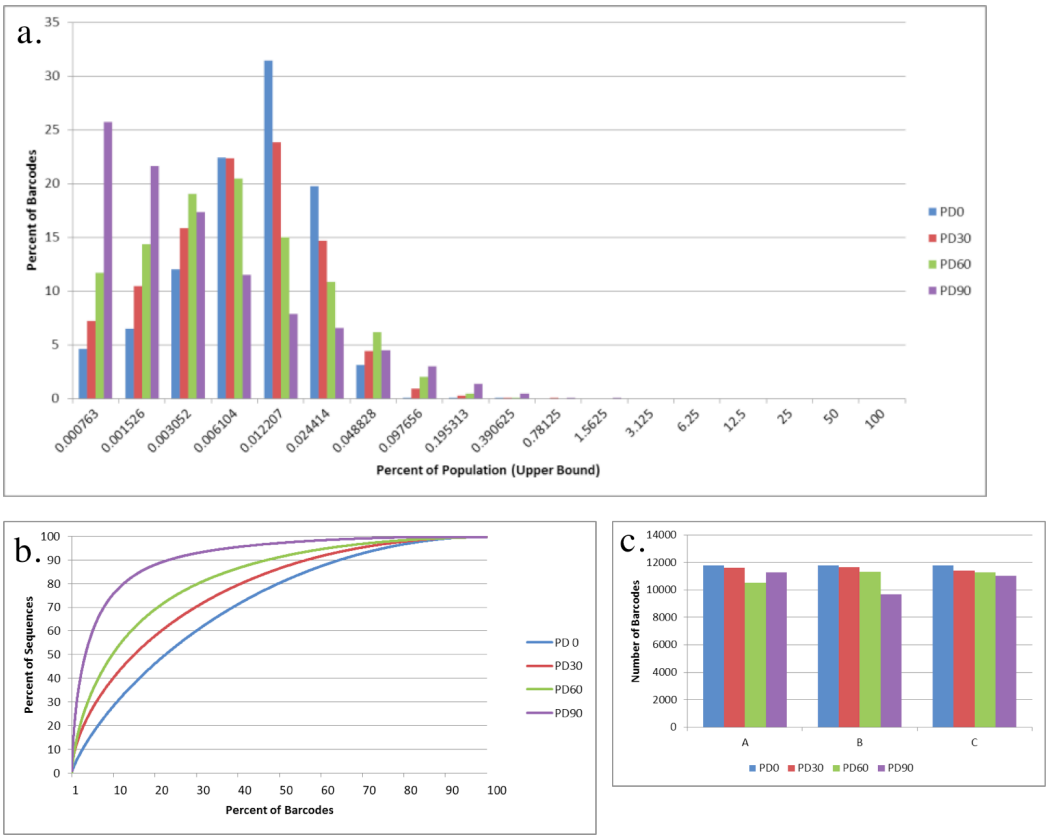


Figure 3.7. Complexity and distribution of clones in HEK-293T barcode experiments.

(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency). (c.) Number of barcodes in each population at each time point.

	Frequency (Count/3x10^5)	
Pop. Doublings	**Median**	**Average**
0	0.006755 (20.3)	0.008472 (25.4)
30	0.005180 (15.5)	0.008565 (25.7)
60	0.003563 (10.7)	0.008777 (26.3)
90	0.001655 (5.0)	0.010305 (30.9)

Table 3.2. Median and average barcode frequency over time in HEK-293T cells.

The median and average frequency (percent of population represented by each barcode), for the population at each timepoint was calculated. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.

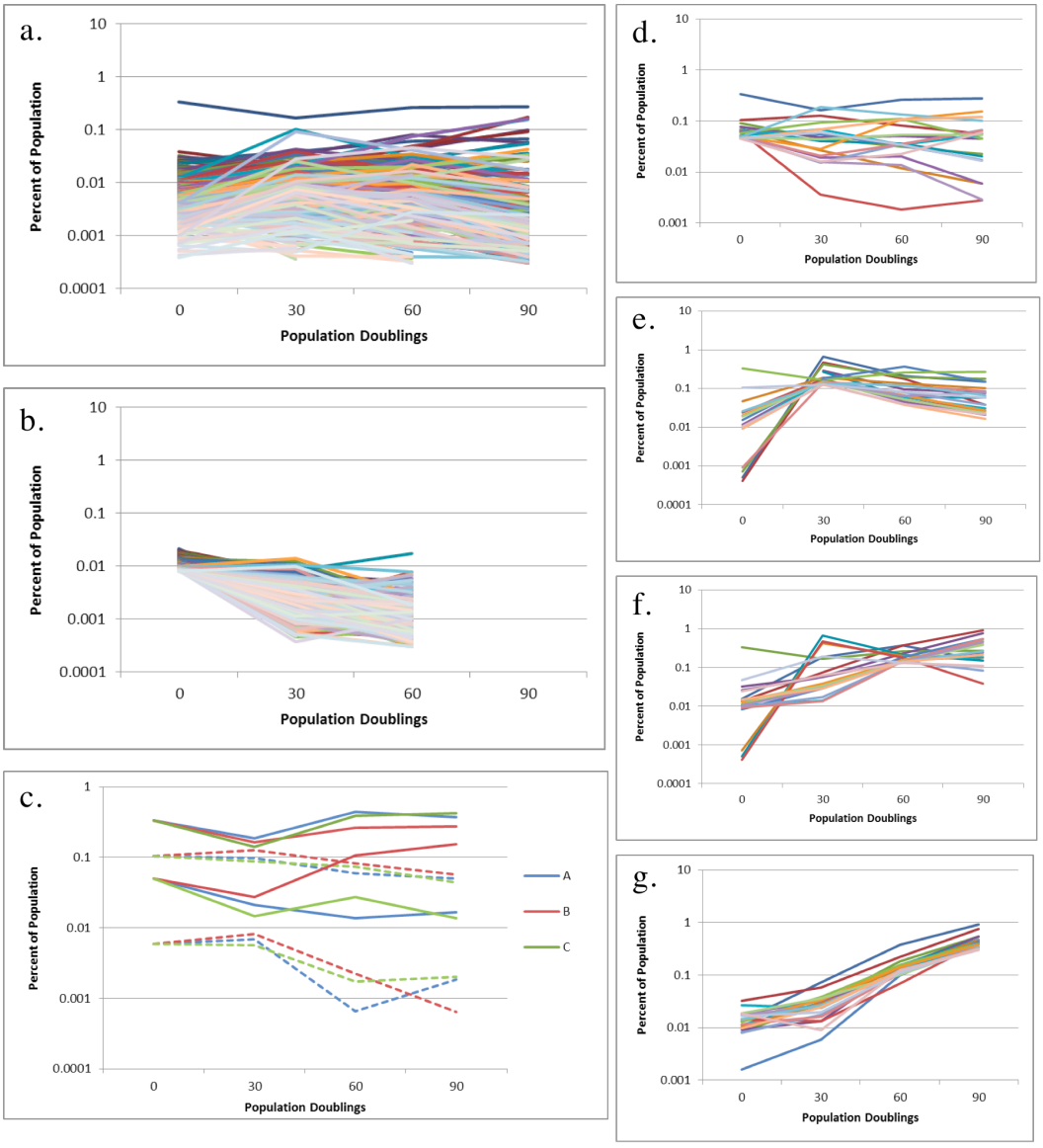

a.
b.
c.
d.
e.
f.
g.
Percent of Population
Population Doublings
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
A
B
C

Figure 3.8. Clonal dynamics of HEK-293T cells.

Each line represents the frequency of a single barcode-marked clone over the course of the experiment. (a) a sampling of the behavior of 250 clones from a cross-section of the population, shown on a log scale. (b) barcodes that disappeared from the population after 30 or 60 population doublings. (c) Four randomly selected barcodes are tracked across all three populations (A, blue; B, red; C, green). (d) The top 20 most frequent barcodes at PD0. (e) The top 20 most frequent barcodes at PD30. (f) The top 20 most frequent barcodes at PD60. (g) The top 20 most frequent barcodes at PD90.

K562 CELLS

Introduction

The K562 cell line was isolated from the pleural effusion of a chronic myelogenous leukemia patient in blast crisis in 19, and was the first immortalized myelogeous leukemia line to be established (Klein, 1976). The cells were originally positive for the classic CML t(9:22) which results in the BCR-Abl1 fusion gene; however, the cells used for this study no longer harbor that marker chromosome (Appendix B), but do have large amplifications of both BCR, Abl1, and the fusion genes (Appendix C).

Materials and Methods

Creation of K562 Barcode Library

K562 cells were obtained from ATCC (Logan, UT). Log-phase cells were spun down and resuspended in a small volume of media (>10^6 cells per mL), and 2 x 10^6 cells were infected with barcode lentivirus calculated to achieve an MOI of 0.05. Cells were diluted in additional media and plated 1 hour after the start of infection. Transduction levels were confirmed by flow cytometry after 48 hours. 4 days after infection, 2 x 10^5 GFP positive cells were isolated by flow sorting on a FACS Aria II (Becton Dickinson). This sorted population of barcoded cells were then expanded in culture for 5 days, then divided for various uses as population doubling zero (PD0) cells.

In vitro culture experiments

K562 cells were obtained from ATCC (Logan, UT). The cells were maintained in RPMI (Cellgro) supplemented with 10% bovine growth serum (Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. 3 aliquots of 300,000 PD0 cells each were separated into individual 6-well tissue culture plates as populations A, B, and C. Every three days, cells were counted and analyzed for GFP expression by flow cytometry. 300,000 cells from each population were transferred to a new well with 3mL of fresh media.

Sequencing sample preparation

Genomic DNA was harvested from pelleted cells with the DNeasy Blood and Tissue Kit (QIAGEN). PCR amplification of barcodes was performed as for the plasmid barcode sequencing in Chapter 2, except that the genomic DNA from 300,000 cells (approximately 5.6ug due to the aneuploidy of K562 cells, Appendix B) was used as template starting material, divided between 8 50uL PCR reactions, which were subsequently pooled.

Results

Barcoded K562 cells grew rapidly, the population doubling a approximately every 23 hours, and the doubling rates of the three populations were similar (Figure 3.9). K562 sequencing results were processed as described previously (Chapter 2). The number of barcodes in each population over time (Figure 3.10c) demonstrates a steady, significant loss of clones over the progression of the experiment. In order to visualize the distribution of barcode frequencies within each timepoint, the frequency of each barcode

was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 3.10a) shows that the distribution of barcode frequencies within the PD0 starting population was relatively normal, indicated by the bell-shaped curve, and the distribution of the curve does not change significantly over time, although the shape broadens slightly.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 3.10b). This plot demonstrates that the barcodes in the earliest timepoint (PD0) are relatively evenly distributed (a perfect distribution would follow the line (y=x), but over the course of the experiment, a small fraction of the barcodes become more dominant, and are represented in a larger percent of the sequences from that sample. The median and average percent frequency of barcodes at each timepoint was calculated (Table 3.3), and agrees with the other data showing a progression from evenly distributed to a separation between low-frequency clones and a few more dominant ones. The median barcode frequency decreases with time, while the average barcode frequency increases dramatically, due to the few more dominant clones in the population at those time points.

In order to look at behavior of individual clones, barcode frequency at each timepoint of the experiment was plotted (Figure 3.11). A sampling of every 300th barcode in the population (Figure 3.11a) gives an unbiased cross-section view of the clonal dynamics. The top twenty most frequent barcodes at each timepoint are plotted individually (Figure 3.11d-g) in order to show trends among the dominant clones. A

selection of barcodes which disappeared from the population by PD60 or PD90 were also plotted (Figure 3.11b).

To determine whether the same clone behaved similarly in parallel populations, four randomly selected barcodes with different starting frequencies were tracked across each of the three populations (Figure 3.11c). These plots show that the clones generally behaved similarly in the parallel populations, with the smallest clone becoming extinct in all three populations.

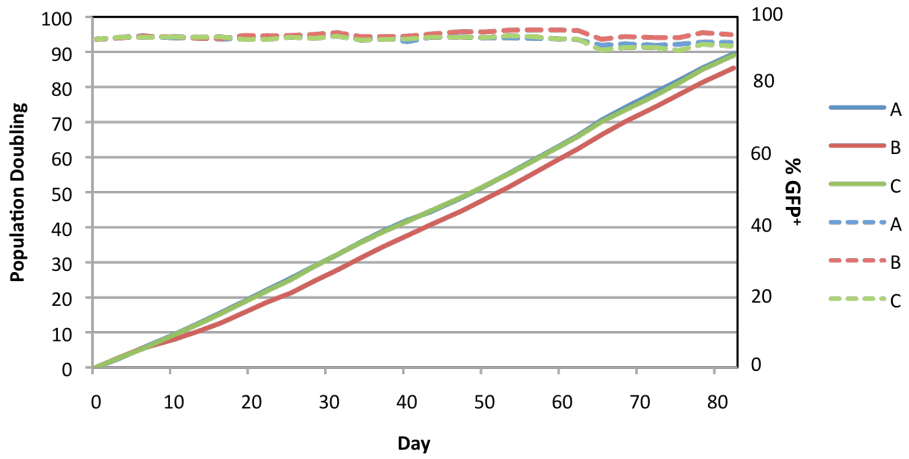


Figure 3.9. Growth and GFP levels of K562 barcode library.

Population doublings of the K562 barcode cells (solid lines) and the percent of the population expressing GFP (dashed lines), over the duration of the experiment.

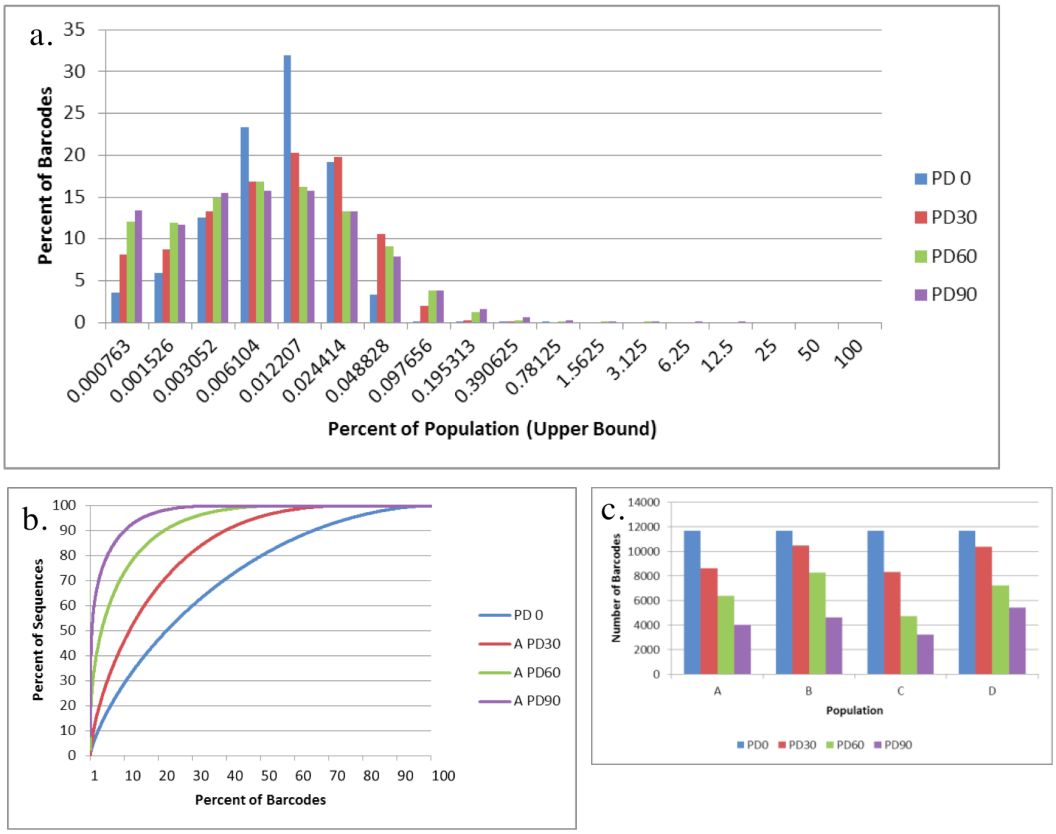


Figure 3.10. Complexity and distribution of clones in K562 experiments.

(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency). (c.) Number of barcodes in each population at each timepoint.

	Frequency (Count/3x10^5)	
Pop. Doublings	**Median**	**Average**
0	0.006755 (20.3)	0.008558 (25.7)
30	0.006835 (20.5)	0.011604 (34.8)
60	0.004709 (14.1)	0.015709 (47.1)
90	0.004594 (13.8)	0.024894 (74.7)

Table 3.3. Median and average barcode frequency over time in K562 cells.

The median and average frequency (percent of population represented by each barcode), for the population at each timepoint was calculated. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.

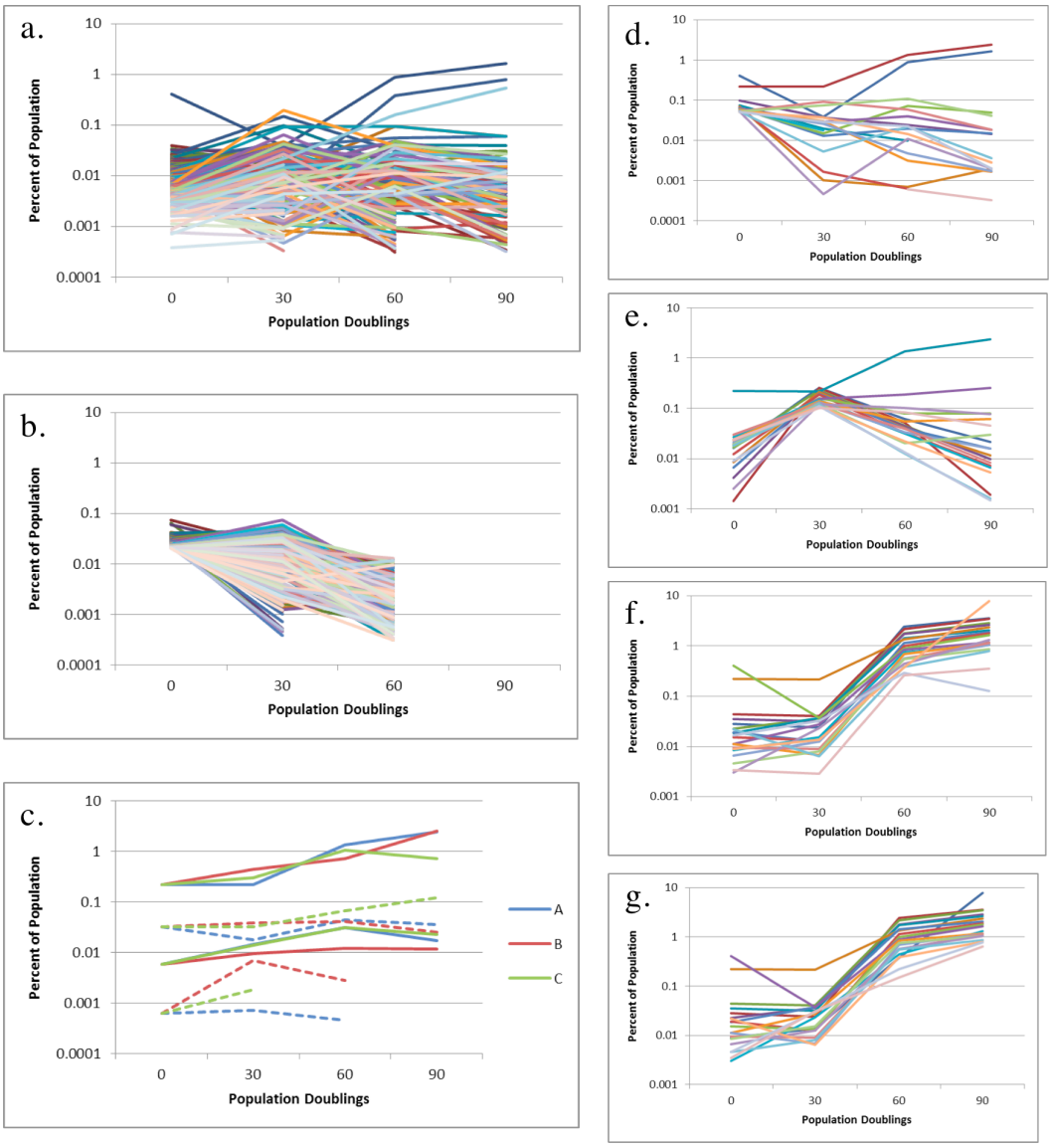

a.
b.
c.
d.
e.
f.
g.
Percent of Population
Population Doublings
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
A
B
C

Figure 3.11. Clonal dynamics of K562 cells.

Each line represents the frequency of a single barcode-marked clone over the course of the experiment. (a) a sampling of the behavior of 250 clones from a cross-section of the population, shown on a log scale. (b) barcodes that disappeared from the population after 30 or 60 population doublings. (c) Four randomly selected barcodes are tracked across all three populations (A, blue; B, red; C, green). (d) The top 20 most frequent barcodes at PD0. (e) The top 20 most frequent barcodes at PD30. (f) The top 20 most frequent barcodes at PD60. (g) The top 20 most frequent barcodes at PD90.

K562 SUBCLONE DERIVED FROM A SINGLE CELL

Introduction

The unexpectedly rapid clonal evolution seen in K562s, a cell line which has been grown in culture for many years, led us to wonder if these clonal differences were due to pre-existing differences within the cell line population, or if these differences were the result of changes the cells were acquiring during the relatively short time of the experiment itself. In order to distinguish between these two possibilities, we sought to study the dynamics and evolution of a population of cells that were as closely related as possible. We started with a single K562 cell, allowed it to expand to suitable numbers (approximately 21 generations), and barcoded and passaged the cells in the same way as for the original K562 barcoded library.

Materials and Methods

A clonal line of K562 cells for these experiments was created by isolating single cells by limiting dilution into a 96 well plate. A single well was expanded to a few million cells, which were then used for barcoding. K562 cells were maintained in RPMI (Cellgro) supplemented with 10% bovine growth serum(Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. 3 aliquots of 300,000 PD0 cells each were separated into individual 6-well tissue culture plates as populations A, B, and C. Every three days, cells were counted and analyzed for GFP expression by flow cytometry. 300,000 cells from each population were transferred to a new well with 3mL of fresh media.

Sequencing sample preparation

Genomic DNA was harvested from pelleted cells with the DNeasy Blood and Tissue Kit (QIAGEN). PCR amplification of barcodes was performed as for the plasmid barcode sequencing in Chapter 2, except that the genomic DNA from 300,000 cells (approximately 5.6ug due to the aneuploidy of K562 cells, Appendix B) was used as template starting material, divided between 8 50uL PCR reactions, which were subsequently pooled.

Results

As with the original barcoded K562 cells, the subcloned cells grew rapidly, the population doubling a approximately every 23 hours, and the doubling rates of the three populations were very similar (Figure 3.12). K562 sequencing results were processed as described previously (Chapter 2). The number of barcodes in each population over time (Figure 3.13c) demonstrates a steady loss of clones over the progression of the experiment, although fewer clones were lost in these experiment than in those with the original K562 library. In order to visualize the distribution of barcode frequencies within each timepoint, the frequency of each barcode was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 3.13a) shows that the distribution of barcode frequencies within the PD0 starting population was relatively normal, indicated by the bell-shaped curve, and the distribution of the curve does not change significantly over time.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 3.13b). This plot demonstrates that the barcodes in the earliest timepoint (PD0) are relatively evenly distributed (a perfect distribution would follow the line (y=x), and over the course of the experiment, the population skews only a fraction of the amount seen in the original K562 library. The median and average percent frequency of barcodes at each timepoint were calculated (Table 3.4), and shows that the population changes very little over the 90 population doublings of the experiment.

In order to look at behavior of individual clones, barcode frequency at each timepoint of the experiment was plotted (Figure 3.14). A sampling of every 300th barcode in the population (Figure 3.14a) gives an unbiased cross-section view of the clonal dynamics. The top twenty most frequent barcodes at each timepoint are plotted individually (Figure 3.14d-g) in order to show trends among the dominant clones. A selection of barcodes which disappeared from the population by PD60 or PD90 were also plotted (Figure 3.14b).

To determine whether the same clone behaved similarly in parallel populations, three randomly selected barcodes with different starting frequencies were tracked across each of the three populations (Figure 3.14c). These plots show that the clones in separate populations generally behaved very similarly over the course of the experiment.

In order to visually compare the original K562 populations to the subcloned K562 barcode library populations, area-proportional Venn diagrams were constructed. These diagrams show the overlap of barcodes between the three parallel populations at each timepoint in the experiment. It is immediately obvious that the clones in the

barcode library created in a population of K562 cells diverged much more rapidly and dynamically than did those of the much more closely related subclonal K562s.

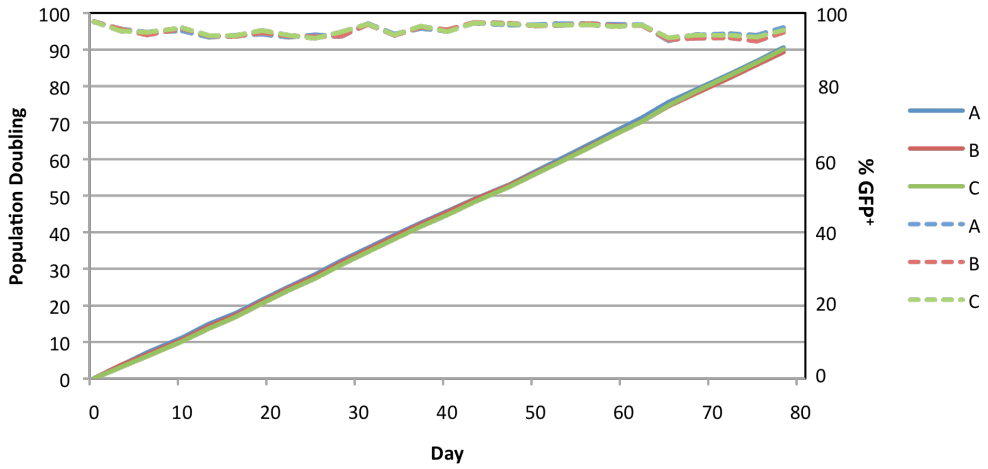


Figure 3.12. Growth and GFP levels of subcloned K562 barcode library.

Population doublings of the subcloned K562 barcode cells (solid lines) and the percent of the population expressing GFP (dashed lines), over the duration of the experiment.

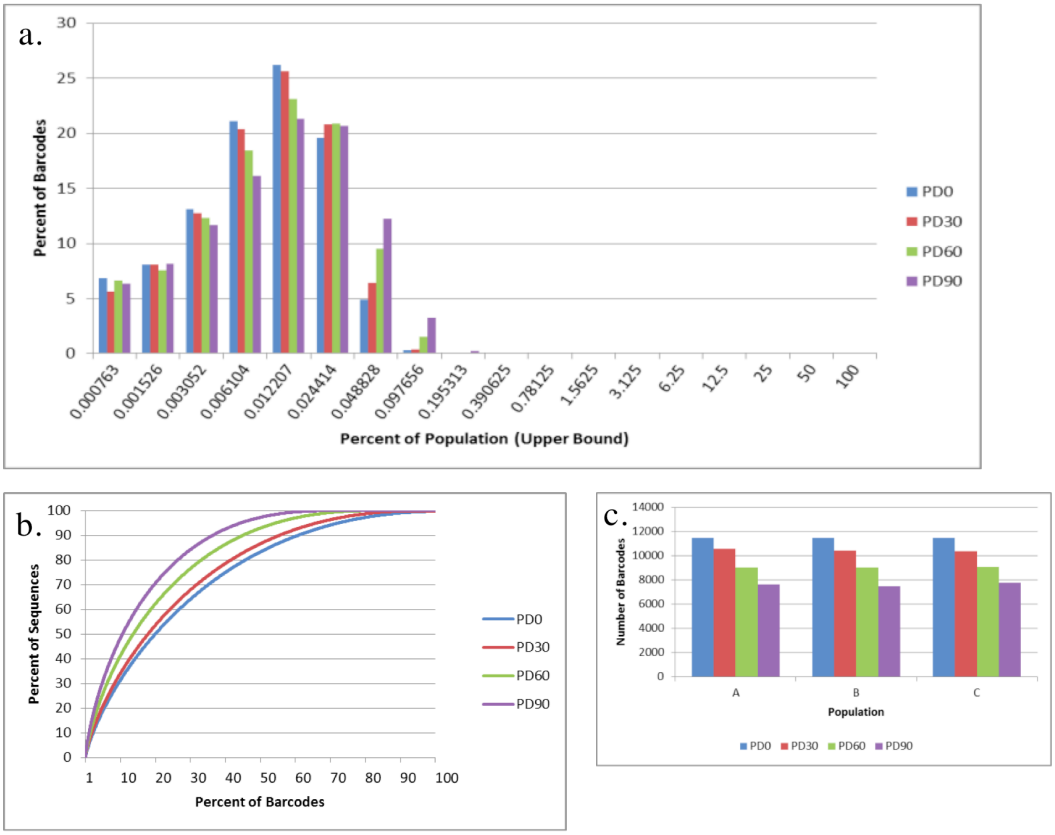


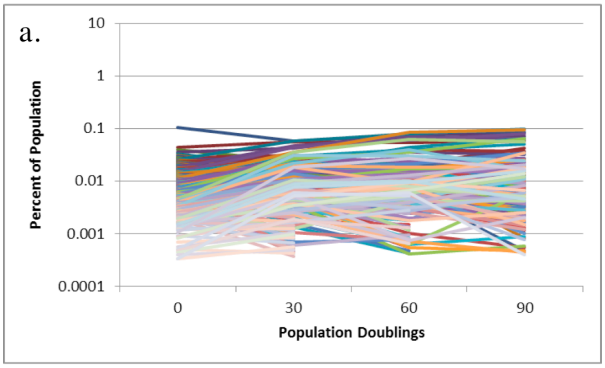
Figure 3.13. Complexity and distribution of clones in subcloned K562 barcode experiments.

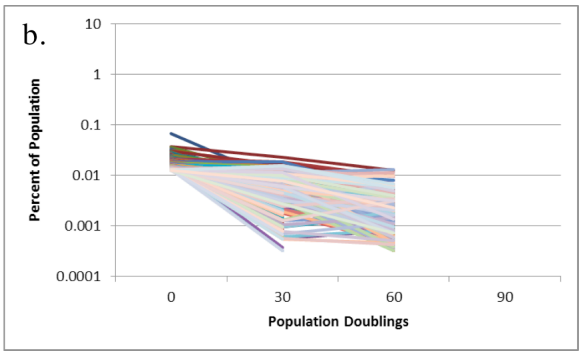
(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency). (c.) Number of barcodes in each population at each timepoint.

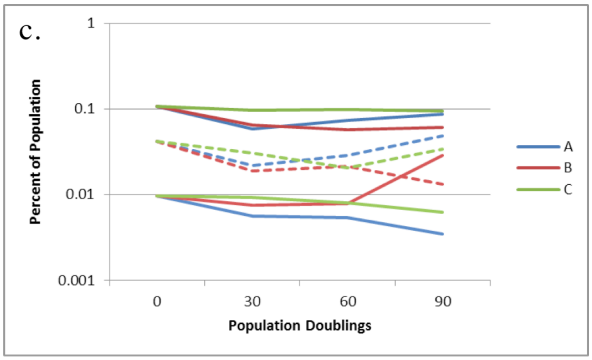
	Frequency (Count/3x10^5)	
Pop. Doublings	**Median**	**Average**
0	0.006278 (18.8)	0.008687 (26.1)
30	0.006740 (20.2)	0.009454 (28.4)
60	0.007208 (21.6)	0.011065 (33.2)
90	0.007901 (23.7)	0.013103 (39.3)

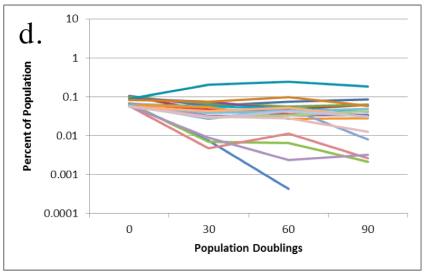
Table 3.4. Median and average frequencies of K562 cells at each timepoint.

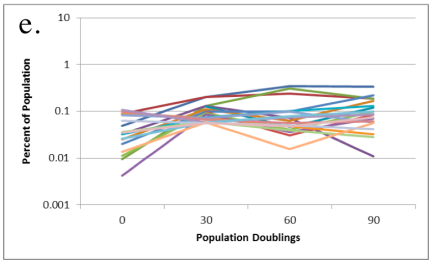
The median and average frequency (percent of population represented by each barcode), for the population at each timepoint was calculated. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.

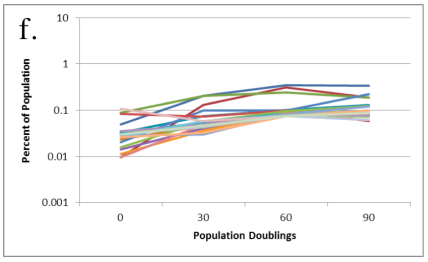

a.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
Population Doublings


b.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
Population Doublings


c.
Percent of Population
1
0.1
0.01
0.001
0
30
60
90
Population Doublings
A
B
C


d.
Percent of Population
10
1
0.1
0.01
0.001
0.0001
0
30
60
90
Population Doublings


e.
Percent of Population
10
1
0.1
0.01
0.001
0
30
60
90
Population Doublings


f.
Percent of Population
10
1
0.1
0.01
0.001
0
30
60
90
Population Doublings

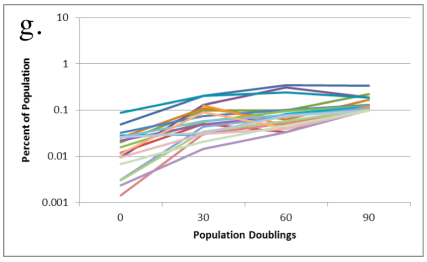

g.
Percent of Population
10
1
0.1
0.01
0.001
0
30
60
90
Population Doublings

Figure 3.14. Clonal dynamics of subcloned K562 cells.

Each line represents the frequency of a single barcode-marked clone over the course of the experiment. (a) a sampling of the behavior of 250 clones from a cross-section of the population, shown on a log scale. (b) barcodes that disappeared from the population after 30 or 60 population doublings. (c) Three randomly selected barcodes are tracked across all three populations (A, blue; B, red; C, green). (d) The top 20 most frequent barcodes at PD0. (e) The top 20 most frequent barcodes at PD30. (f) The top 20 most frequent barcodes at PD60. (g) The top 20 most frequent barcodes at PD90.

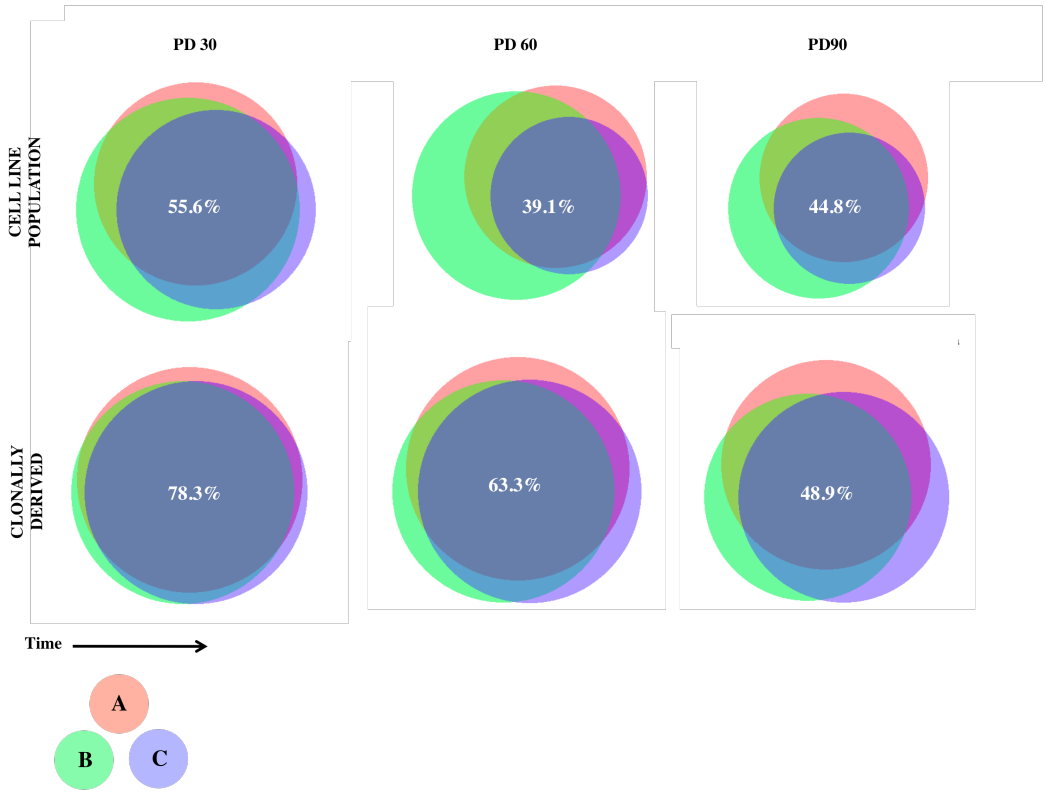


Figure 3.15. Comparison of original and subcloned K562 barcode libraries.

The three parallel populations grown from the same PD0 cells from both K562 barcode libraries were compared to one another at each timepoint and displayed as area proportional Venn diagrams. The percent of barcodes found in all of the three populations is indicated. Overlap of barcodes between any two of the populations can been seen, as well as the relative proportion of barcodes which were unique to that population. A, red; B, green; and C, blue. All diagrams were made with the BioVenn web application which can be found at www.cmbi.ru.nl/cdd/biovenn/ (Hulsen, 2008).

CHAPTER FOUR
Comparison of clonal dynamics observed in vitro and in vivo

HCC827 CELLS CULTURED IN VITRO

Introduction

In order to compare the clonal dynamics of a cell line in vivo as well as in vitro, we chose to apply our barcode system to HCC827 cells. HCC827 is a human cancer cell line derived from a non-small cell lung cancer by Adi Gazdar and John Minna in 1994. HCC827s have been shown to readily form tumors in nude mice, making them an ideal candidate for our experiments. In order to be able to compare the cells grown in vitro as well as in the mice, we used aliquots of the same PD0 population as our experimental starting point (Figure 4.1).

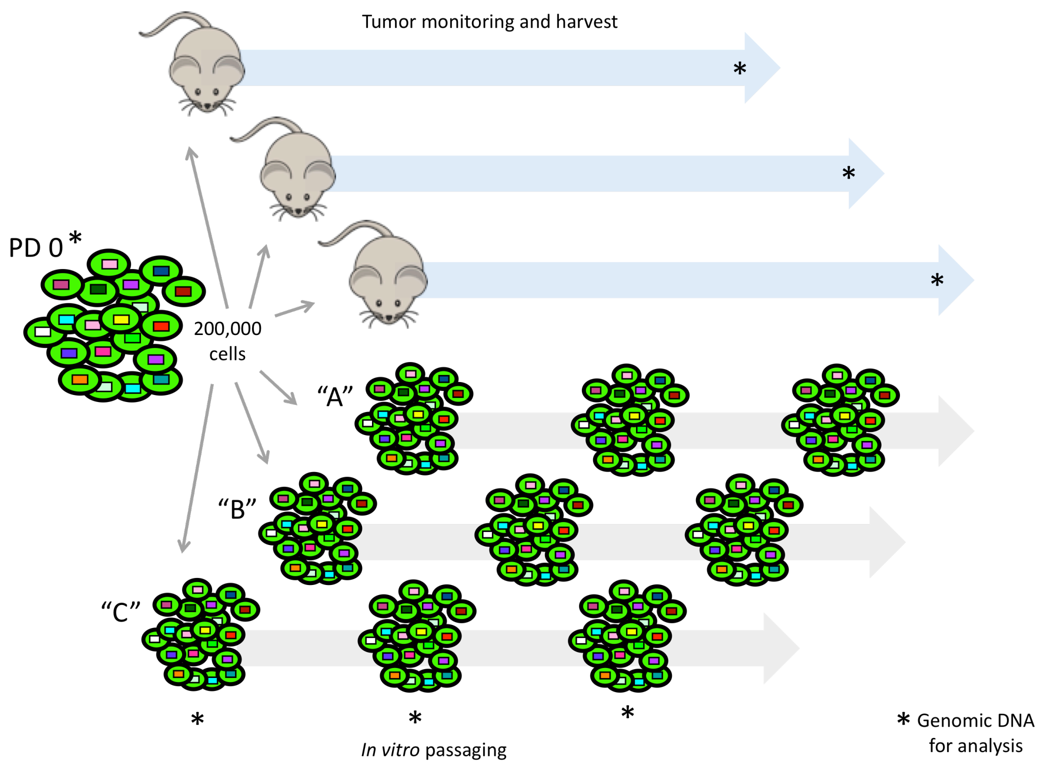


Figure 4.1. Diagram of Experimental Design for study of HCC827 cell barcode library in vitro and in vivo.

Materials and Methods

Creation of HCC827 Barcode Library

HCC827 cells were obtained from ATCC (Logan, UT). Log-phase cells were trypsinized, resuspended in a small volume of media (>10^6 cells per mL), and 2 x 10^6 cells were infected with barcode lentivirus calculated to achieve an MOI of 0.05. Cells were diluted in additional media and plated 1 hour after the start of infection. Transduction levels were confirmed by flow cytometry after 48 hours. 7 days after infection, 2 x 10^5 GFP positive cells were isolated by flow sorting on a FACS Aria II (Becton Dickinson). This sorted population of barcoded cells were then expanded in culture for 7 days, then divided for various uses as population doubling zero (PD0) cells.

In vitro culture experiments

HCC827 cells were maintained in RPMI (Cellgro) supplemented with 10% bovine growth serum(Hyclone), 100 units/mL penicillin, 100ug/mL streptomycin, and 2mM L-glutamine. 3 aliquots of 300,000 PD0 cells each were separated into individual 6-well tissue culture plates as populations A, B, and C. Every three days, cells were trypsinized, counted, and analyzed for GFP expression by flow cytometry. 200,000 cells from each population were transferred to a new well with 3mL of fresh media.

Sequencing sample preparation

Genomic DNA was harvested from pelleted cells with the DNeasy Blood and Tissue Kit (QIAGEN). PCR amplification of barcodes was performed as for the plasmid barcode sequencing in Chapter 2, except that the genomic DNA from 200,000 cells

(approximately 2.5ug due to the aneuploidy of HCC827 cells, Appendix D) was used as template starting material, divided between 8 50uL PCR reactions, which were subsequently pooled.

Results

Barcoded HCC827 cells grew very slowly in culture, the population doubling approximately every 3 days. (Figure 4.2). HCC827 sequencing results were processed as described previously (Chapter 2). In order to visualize the distribution of barcode frequencies within each timepoint, the frequency of each barcode was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 4.3a) shows that the distribution of barcode frequencies within the PD0 starting population was not normal, and over time most barcodes became less frequent as a few became overrepresented, as seen by the shift of barcodes in the PD30 sample to the lower frequency bins, while a single barcode begin to take over the population.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 3.7b). Again, this plot demonstrates that the barcodes in the earliest timepoint (PD0) were significantly skewed early on, and this skew continued to increase with time as a single barcode came to represent more and more of the population. The median and average percent frequency of barcodes at each timepoint was calculated (Table 4.1), and agrees with the other data showing a

progressively more skewed population. The median barcode frequency greatly decreases with time, while the average barcode frequency increases, representative of this single, dominant clone.

In order to look at behavior of individual clones, barcode frequency at each timepoint of the experiment was plotted (Figure 4.4). A sampling of every 300th barcode in the population (Figure 4.4a) gives an unbiased cross-section view of the clonal dynamics. The top twenty most frequent barcodes at each timepoint are plotted individually (Figure 4.4c-f) in order to show trends among the dominant clones.

To determine whether the same clone behaved similarly in parallel populations, three randomly selected barcodes with different starting frequencies were tracked across each of the three populations (Figure 4.4b). These plots show the dominant clone's progress in each of the three populations.

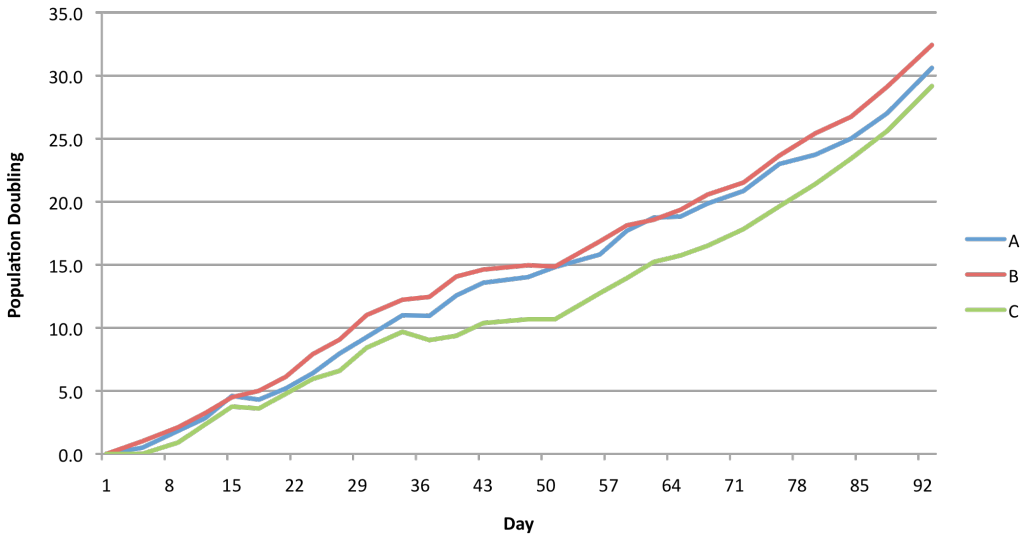
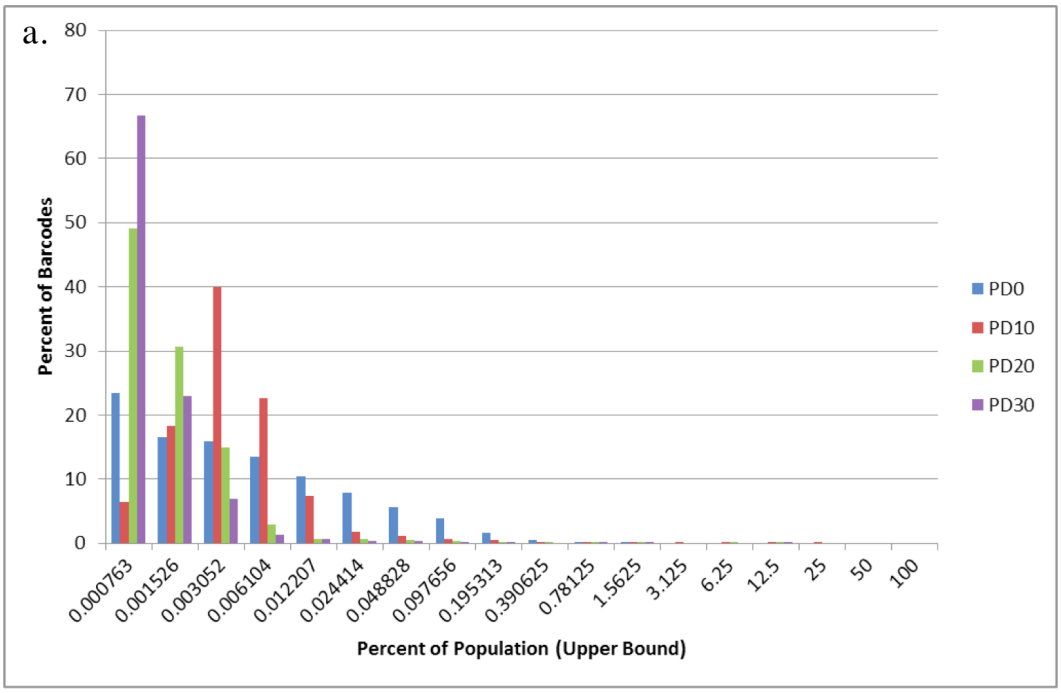


Figure 4.2. Growth of HCC827 barcode cells in vitro.

Population doublings of the HCC827 in vitro barcode cells were plotted as a function of time.


a.
80
70
60
50
40
30
20
10
0
Percent of Barcodes
0.000763
0.001526
0.003052
0.006104
0.012207
0.024414
0.048828
0.097656
0.195313
0.390625
0.78125
1.5625
3.125
6.25
12.5
25
50
100
Percent of Population (Upper Bound)
PD0
PD10
PD20
PD30

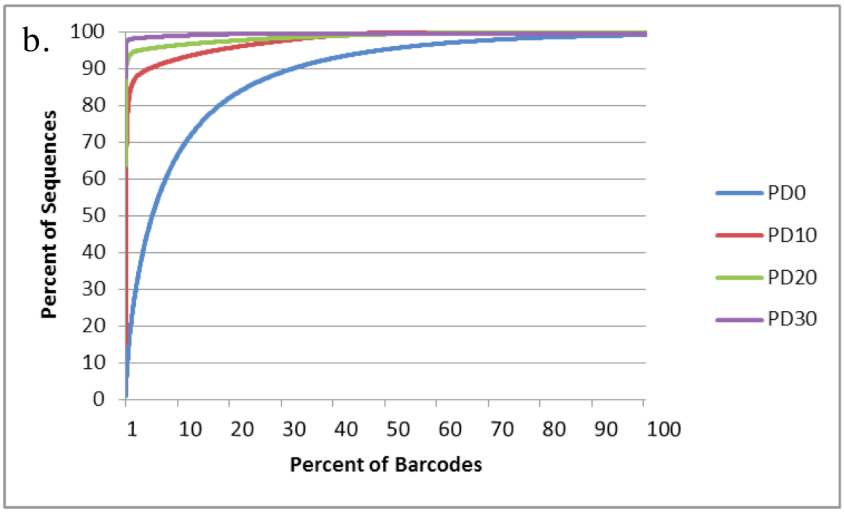

b.
100
90
80
70
60
50
40
30
20
10
0
Percent of Sequences
1
10
20
30
40
50
60
70
80
90
100
Percent of Barcodes
PD0
PD10
PD20
PD30

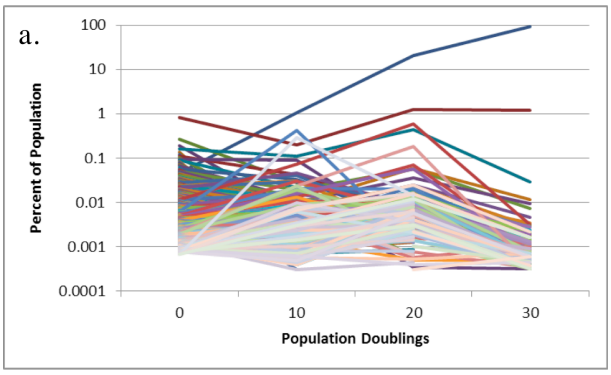
Figure 4.3. Complexity and distribution of clones in HCC827 barcode library in vitro.

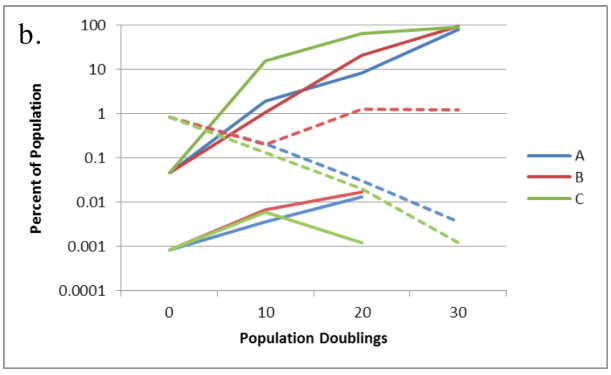
(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency).

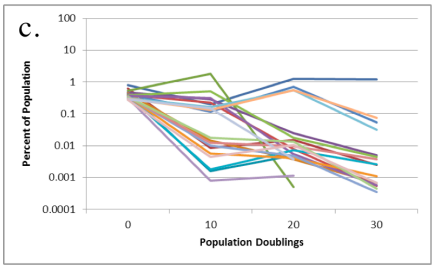
	Frequency (Count/3x10^5)	
Pop. Doublings	**Median**	**Average**
0	0.002354 (7.1)	0.012422 (37.3)
30	0.001664 (5.0)	0.010586 (31.8)
60	0.003032 (9.1)	0.008797 (26.4)
90	0.000738 (2.2)	0.027162 (81.5)

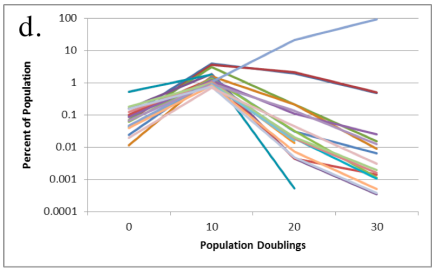
Table 4.1. Median and average barcode frequency over time in HCC827 cells.

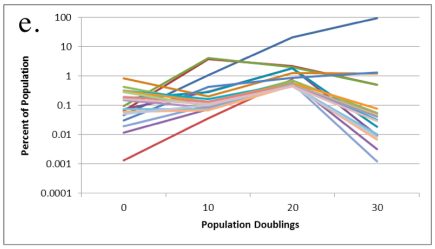
The median and average frequency (percent of population represented by each barcode), for the population at each timepoint was calculated. The number of times a barcode with the indicated frequency would be present in a population of $3x10^5$ cells is listed in parenthesis.


a.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings


b.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings
A
B
C


c.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings


d.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings


e.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings

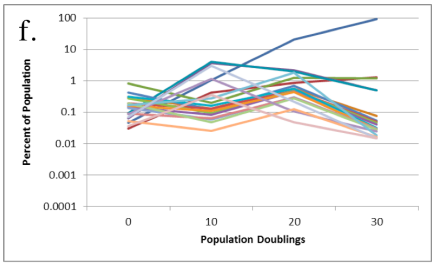

f.
Percent of Population
100
10
1
0.1
0.01
0.001
0.0001
0
10
20
30
Population Doublings

Figure 4.4. Clonal dynamics of HCC827 cells in vitro.

Each line represents the frequency of a single barcode-marked clone over the course of the experiment. (a) a sampling of the behavior of 250 clones from a cross-section of the population, shown on a log scale. (b) barcodes that disappeared from the population after 30 or 60 population doublings. (c) Three randomly selected barcodes are tracked across all three populations (A, blue; B, red; C, green). (d) The top 20 most frequent barcodes at PD0. (e) The top 20 most frequent barcodes at PD30. (f) The top 20 most frequent barcodes at PD60. (g) The top 20 most frequent barcodes at PD90.

HCC827 CELLS CULTURED IN VIVO

Materials and Methods

HCC827 barcode library cells, "PD 0" were trypsinized and resuspended at $2x10^6$ cells per mL. $2x10^5$ cells (100uL) were injected subcutaneously into the left flank of each of five 8 week old homozygous nude female mice (Charles River). Mice were monitored weekly for tumor appearance, and then tumors were measured twice weekly with digital calipers and tumor volumes were calculated with the formula (1/2 width by width by length). Mice were sacrificed when tumors reached 1 cm^3 or at the experiment termination point. Four mice had tumors, three of which were large enough to obtain genomic DNA and sequence (DNeasy Blood and Tissue Kit, QIAGEN). No metastases were seen upon dissection.

Results

Tumor sizes were somewhat unstable, which may represent the error rate of the measurements or perhaps different levels of inflammation of the tumor site at different timepoints. HCC827 sequencing results were processed as described previously (Chapter 2). In order to compare the distribution of barcode frequencies between tumors, the frequency of each barcode was determined, and then binned logarithmically in order to cluster barcodes with similar frequencies. The results were normalized to account for different numbers of barcodes in each sample. This histogram (Figure 4.5a) shows that the distribution of barcode frequencies was very similar among the three tumors.

Another way of visualizing the distribution of the sequences among the barcodes in a population is shown by plotting the percent of the sequences that are taken up by what percent of the barcodes (Figure 4.5b). Again, this plot demonstrates that the distribution of barcodes within each of the tumors is very similar, and there were similar numbers of barcodes in each sample (Figure 4.5c). Finally, an area-proportional Venn diagram was generated (Figure 4.5d), showing that a majority of the barcodes were found in all three tumors.

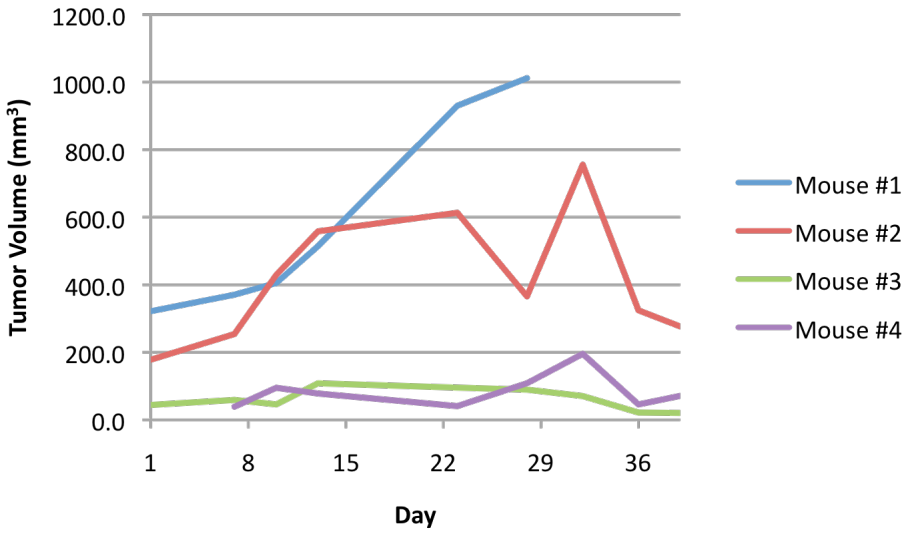
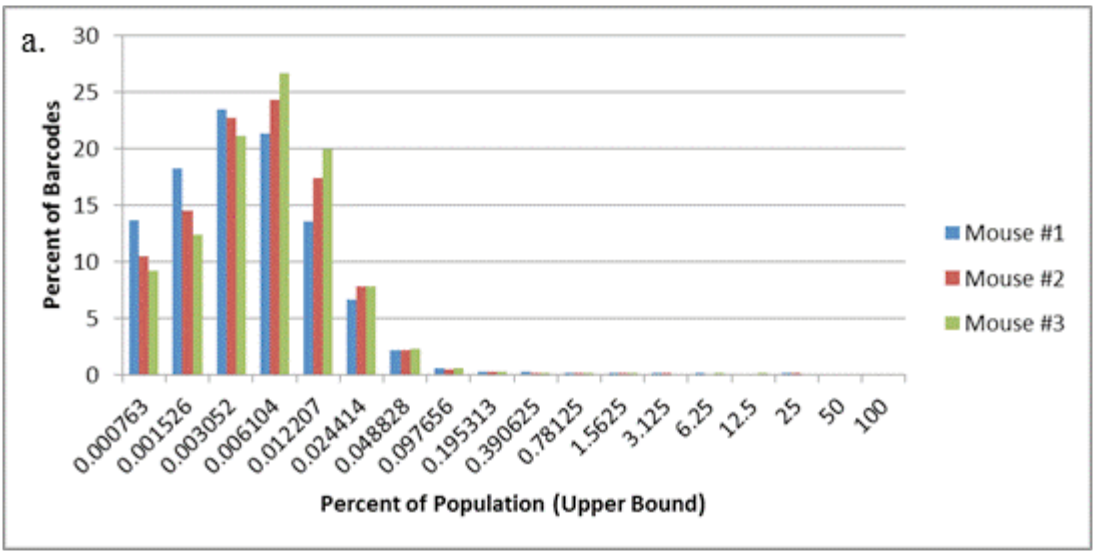
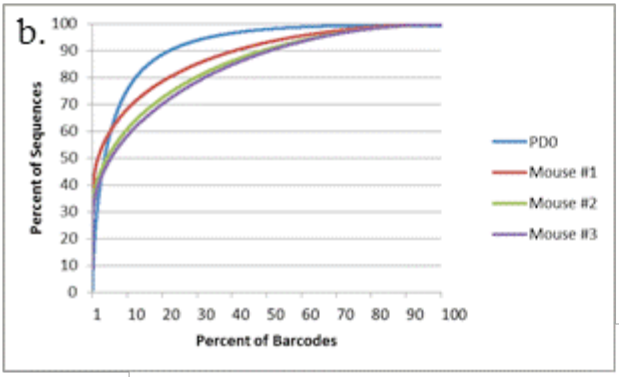
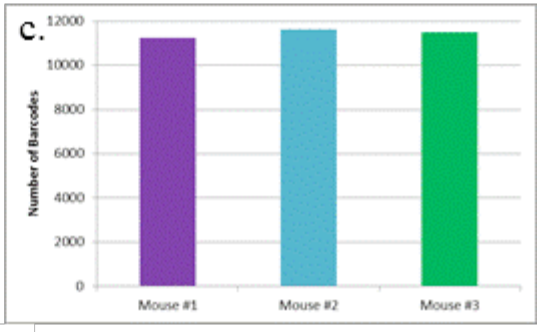


Figure 4.5. HCC827 xenograft tumor growth.

Mice were monitored twice weekly and tumor sizes were measured with digital calipers.


a.
Percent of Barcodes
30
25
20
15
10
5
0
0.000763
0.001526
0.003052
0.006104
0.012207
0.024414
0.048828
0.097656
0.195313
0.390625
0.78125
1.5625
3.125
6.25
12.5
25
50
100
Percent of Population (Upper Bound)
Mouse #1
Mouse #2
Mouse #3


b.
Percent of Sequences
100
90
80
70
60
50
40
30
20
10
0
1
10
20
30
40
50
60
70
80
90
100
Percent of Barcodes
PD0
Mouse #1
Mouse #2
Mouse #3


c.
Number of Barcodes
12000
10000
8000
6000
4000
2000
0
Mouse #1
Mouse #2
Mouse #3

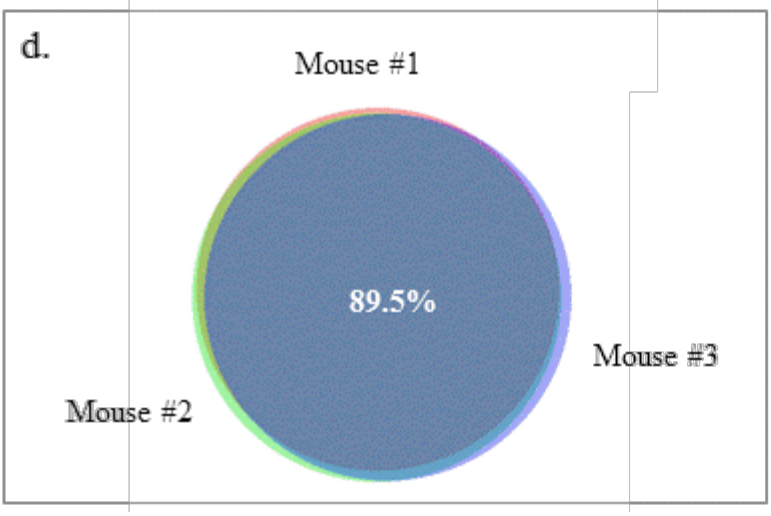

d.
Mouse #1
89.5%
Mouse #3
Mouse #2

Figure 4.6. Complexity and distribution of HCC827 barcode library xenograft tumors.

(a.) Histogram of the distribution of barcode frequencies. Barcodes were grouped by frequency into log 2 scale bins and plotted as the percent of barcodes within each bin's range of frequencies. (b.) Linear curve plot displaying the percent of the sequences made up of what percent of the barcodes (in decreasing order of frequency). (c.) Number of barcodes in each population at each timepoint. (d.) Venn diagram (BioVenn, http://www.cmbi.ru.nl/cdd/biovenn/) with areas proportional to the overlap of barcodes sequenced from the 3 mouse tumors. 89.5% of barcodes were found in all three tumors.

CHAPTER FIVE
Conclusions and Future Directions

CONCLUSIONS

The data presented here demonstrate that we were successful in creating a sensitive, heritable, and quantitative high-throughput system to track thousands of clones within a population of cells. This system revealed that dynamics and clonal evolution is present in common cell lines after decades of culturing even when cultured under ideal conditions. Comparisons of K562s barcoded either as a population or as a subclone derived from a single cell with a minimal number of doublings between the cells within the population reveal that clonal dynamics is ongoing within clonal populations, but is slower in more closely related populations, as might be expected. The application of the barcoding system to study the dynamics of a cell line in vitro and in vivo demonstrated that at least for the cell line used, in vivo expansion is less selective than in vitro tissue culture, an unexpected result.

Our system has the benefits of both a large complexity and barcoded cell selection, as well as taking advantage of the high throughput advantages of next gen sequencing, targeted barcode marking, and direct PCR-mediated one step sample preparation to reduce handling and potential skewing of the sample after the fact. We also take advantage of the fact that the system is vastly applicable to all kinds of studies, from cancer to gene therapy, stem cells to virology, genetics to tissue regeneration, transplantation, etc. Can use this system to answer a number of previously un-answerable questions. We have validated this system and used to study a number of interesting things: clonal dynamics of cell lines, and cells in vivo vs. in vitro.

FUTURE DIRECTIONS

Improved barcode design

When constructing newer barcode libraries, we have begun to include a third unique restriction site between the two sites that will be used for barcode insertion, to allow for digestion and removal of any remaining undigested or re-circularized vector after ligation of the barcode fragments to remove any unbarcoded vectors prior to lentivirus production.

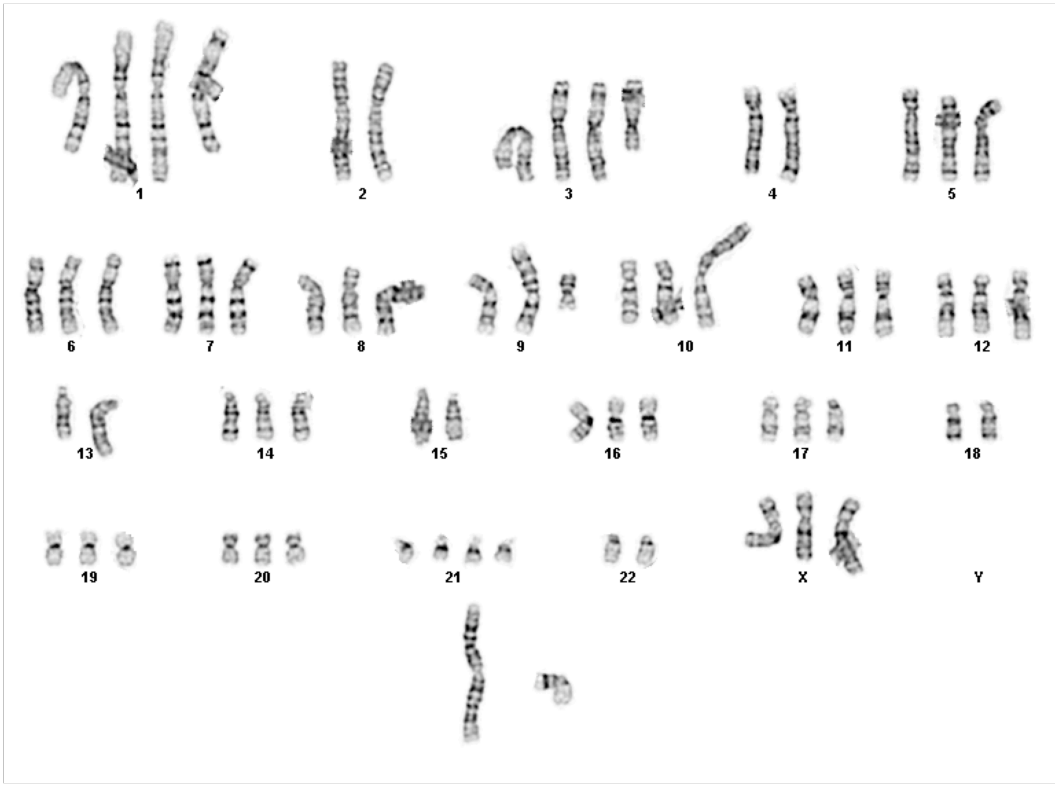
As we found after analyzing the barcode library presented here, it is best to have variable bases hand mixed to equal ratios during oligo synthesis to prevent the unequal base composition often obtained with machine mixed bases

To aid in downstream sequence analysis, it may be useful to limit the size of homopolymer stretches within the barcode to no more than 4. To do this, we suggest the following design: NNNSW NNNSW NNNSW NNNSW, where Ns can be any of the 4 bases, S can be either a C or G, and W can be either an A or T. An additional feature that may aid in quality analysis of the barcodes is to anchor one base toward the end of the barcode. Such as anchoring a T at position 15, thusly: NNNSW NNNSW NNNS**T** NNNSW. It is important to note that these modifications reduce the number of possible barcode sequences to approximately 2.1 billion (4 to the 12^{th} power times 2 to the 7^{th} power).

Applications of the barcode tracking system

The barcode tracking system we have developed is applicable to a broad array of applications. Our lab is already applying the barcode system demonstrated here to a number of additional biological questions. We have created a number of additional barcode libraries in different lentiviral vector backbones with complexities ranging from 30,000 to 1.5 million barcodes. Current or planned studies in our lab include tracking the clonality and dynamics of primary cells grown in vitro, monitoring the dynamics and clonal evolution of patient leukemia samples before and after chemotherapy treatment, and using barcode lentivirus to track the clonality of mouse HSC transplants. In addition, we have a number of collaborators already using the barcode system and the newer, more complex barcode libraries to answer important questions in their own research. Ongoing collaborative studies include applying our barcode technology to study the clonality of primary and metastatic tumors in mice, to sensitively and quantitatively track the clonal dynamics of hematopoietic stem cell transplant and engraftment in primates, and tracking the clonality of T-cell recruitment, among others. A host of additional potential applications for this technology can be imagined.

APPENDIX A
Karyotype of HEK-293T cells

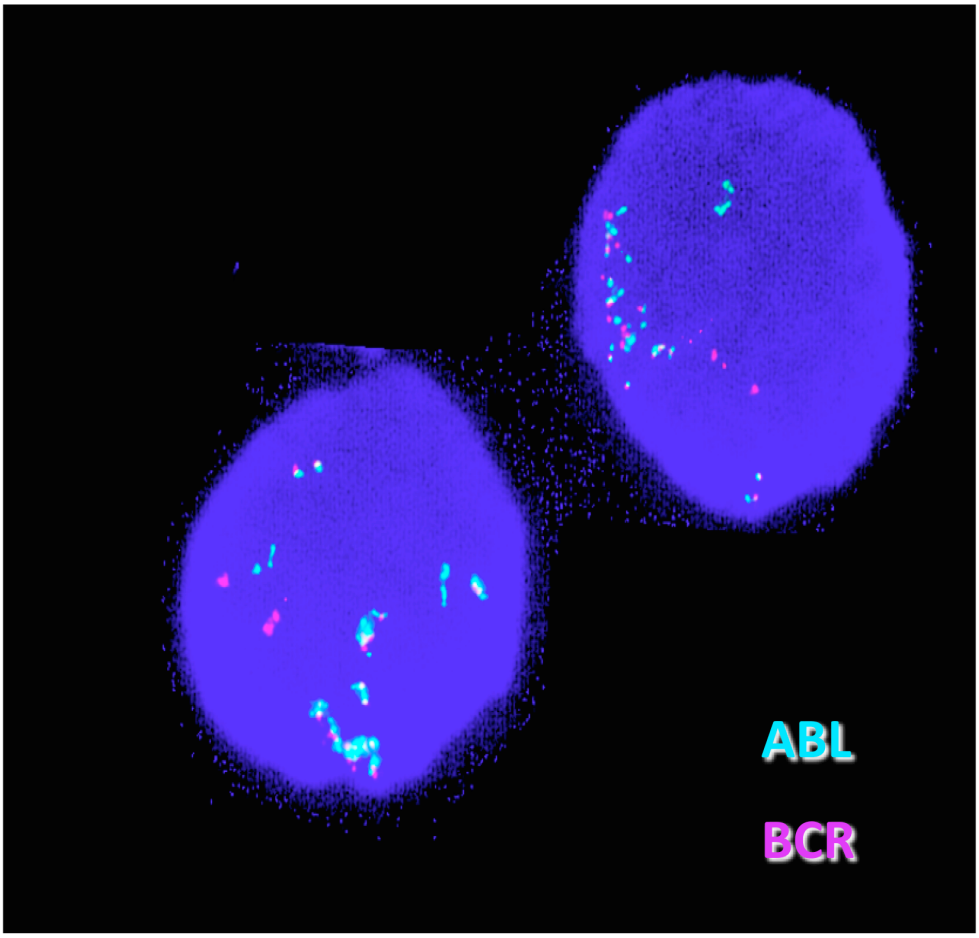


Prepared by Stanford University School of Medicine Cytogenetics Laboratory, 2012.

APPENDIX B
Karyotype of K562 cells

Prepared by Stanford University School of Medicine Cytogenetics Laboratory, 2011

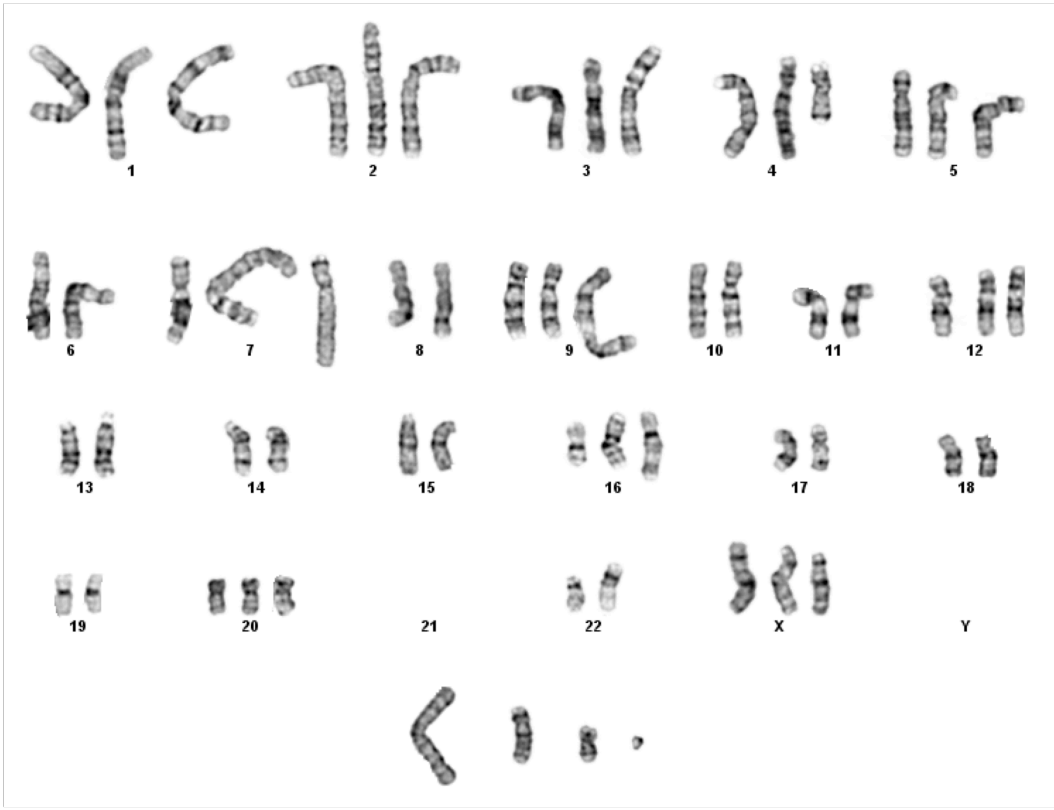
APPENDIX C
BCR-ABL1 FISH of K562 cells



Prepared by Stanford University School of Medicine Cytogenetics Laboratory, 2011

BCR gene amplifications are shown in red, ABL1 genes in blue, and BCR-ABL1 rearrangements are shown in yellow (merge).

APPENDIX D
Karyotype of HCC827 cell line



Prepared by Stanford University School of Medicine Cytogenetics Laboratory, 2011

BIBLIOGRAPHY

Abramson, S. (1977). "The identification in adult bone marrow of pluripotent and restricted stem cells of the myeloid and lymphoid systems." J Exp Med **145**(6): 1567-79.

Adair, J. E. (2012). "Extended survival of glioblastoma patients after chemoprotective HSC gene therapy." Sci Transl Med **4**(133): 133ra57.

Anderson, K. (2011). "Genetic variegation of clonal architecture and propagating cells in leukaemia." Nature **469**(7330): 356-61.

Biasco, L. (2011). "Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell." EMBO Mol Med **3**(2): 89-101.

Capel, B. (1990). "Long- and short-lived murine hematopoietic stem cell clones individually identified with retroviral integration markers." Blood **75**(12): 2267-70.

Carlson, C. A. (2012). "Decoding cell lineage from acquired mutations using arbitrary deep sequencing." Nat Methods **9**(1): 78-80.

Cartier, N. (2009). "Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy." Science **326**(5954): 818-23.

Cavazzana-Calvo, M. (2010). "Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia." Nature **467**(7313): 318-22.

Cepko, C. L. (1998). "Lineage analysis using retroviral vectors." Methods **14**(4): 393-406.

Ciuffi, A. (2009). "Methods for integration site distribution analyses in animal cell genomes." Methods **47**(4): 261-8.

Drize, N. J. (1996). "Local clonal analysis of the hematopoietic system shows that multiple small short-living clones maintain life-long hematopoiesis in reconstituted mice." Blood **88**(8): 2927-38.

Dull, T. (1998). "A third-generation lentivirus vector with a conditional packaging system." J Virol **72**(11): 8463-71.

Dykstra, B. (2007). "Long-term propagation of distinct hematopoietic differentiation programs in vivo." Cell Stem Cell **1**(2): 218-29.

Ellis, B. L. (2012). "Zinc-finger nuclease-mediated gene correction using single AAV vector transduction and enhancement by Food and Drug Administration-approved drugs." Gene Ther.

Ford, C. E. (1956). "Cytological identification of radiation-chimaeras." Nature **177**(4506): 452-4.

Gabriel, R. (2009). "Comprehensive genomic access to vector integration in clinical gene therapy." Nat Med **15**(12): 1431-6.

Gentner, B. (2003). "Rapid detection of retroviral vector integration sites in colony-forming human peripheral blood progenitor cells using PCR with arbitrary primers." Gene Ther **10**(9): 789-94.

Gerrits, A. (2010). "Cellular barcoding tool for clonal analysis in the hematopoietic system." Blood **115**(13): 2610-8.

Gerstung, M. (2012). "Reliable detection of subclonal single-nucleotide variants in tumour cell populations." Nat Commun **3**: 811.

Golden, J. A. (1995). "Construction and characterization of a highly complex retroviral library for lineage analysis." Proc Natl Acad Sci U S A **92**(12): 5704-8.

Hacein-Bey-Abina, S. (2003). "LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1." Science **302**(5644): 415-9.

Harkey, M. A. (2007). "Multiarm high-throughput integration site detection: limitations of LAM-PCR technology and optimization for clonal analysis." Stem Cells Dev **16**(3): 381-92.

Harrison, D. E. (1988). "Number and continuous proliferative pattern of transplanted primitive immunohematopoietic stem cells." Proc Natl Acad Sci U S A **85**(3): 822-6.

Hayakawa, J. (2009). "Long-term vector integration site analysis following retroviral mediated gene transfer to hematopoietic stem cells for the treatment of HIV infection." PLoS One **4**(1): e4211.

Hou, Y. (2012). "Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm." Cell **148**(5): 873-85.

Howe, S. J. (2008). "Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients." J Clin Invest **118**(9): 3143-50.

Hulsen, T. (2008). "BioVenn - a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams." BMC Genomics **9**: 488.

Jordan, C. T. and I. R. Lemischka (1990). "Clonal and systemic analysis of long-term hematopoiesis in the mouse." Genes Dev **4**(2): 220-32.

Klein, E. (1976). "Properties of the K562 cell line, derived from a patient with chronic myeloid leukemia." Int J Cancer **18**(4): 421-31.

Korczak, B. (1988). "Genetic tagging of tumor cells with retrovirus vectors: clonal analysis of tumor growth and metastasis in vivo." Mol Cell Biol **8**(8): 3143-9.

Kustikova, O. (2005). "Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking." Science **308**(5725): 1171-4.

Lu, R. (2011). "Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding." Nat Biotechnol **29**(10): 928-33.

Mazurier, F. (2004). "Lentivector-mediated clonal tracking reveals intrinsic heterogeneity in the human hematopoietic stem cell compartment and culture-induced stem cell impairment." Blood **103**(2): 545-52.

McKenzie, J. L. (2006). "Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment." Nat Immunol **7**(11): 1225-33.

Mitsuhashi, J. (2007). "Retroviral integration site analysis and the fate of transduced clones in an MDR1 gene therapy protocol targeting metastatic breast cancer." Hum Gene Ther **18**(10): 895-906.

Mullighan, C. G. (2011). "Genomic profiling of B-progenitor acute lymphoblastic leukemia." Best Pract Res Clin Haematol **24**(4): 489-503.

Mullighan, C. G. (2008). "Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia." Science **322**(5906): 1377-80.

Naldini, L. (2011). "Ex vivo gene transfer and correction for cell-based therapies." Nat Rev Genet **12**(5): 301-15.

Neil, J. C. and E. R. Cameron (2002). "Retroviral insertion sites and cancer: fountain of all knowledge?" Cancer Cell **2**(4): 253-5.

Nolta, J. A. (1996). "Transduction of pluripotent human hematopoietic stem cells demonstrated by clonal analysis after engraftment in immune-deficient mice." Proc Natl Acad Sci U S A **93**(6): 2414-9.

Pryciak, P. M. and H. E. Varmus (1992). "Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection." Cell **69**(5): 769-80.

Roeder, I. (2008). "Characterization and quantification of clonal heterogeneity among hematopoietic stem cells: a model-based approach." Blood **112**(13): 4874-83.

Satoh, T. and D. M. Fekete (2009). "Lineage analysis of inner ear cells using genomic tags for clonal identification." Methods Mol Biol **493**: 47-63.

Schepers, K. (2008). "Dissecting T cell lineage relationships by cellular barcoding." J Exp Med **205**(10): 2309-18.

Scherer, W. F. (1953). "Studies on the propagation in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix." J Exp Med **97**(5): 695-710.

Schmidt, M. (2003). "Clonality analysis after retroviral-mediated gene transfer to CD34+ cells from the cord blood of ADA-deficient SCID neonates." Nat Med **9**(4): 463-8.

Schmidt, M. (2001a). "A model for the detection of clonality in marked hematopoietic stem cells." Ann N Y Acad Sci **938**: 146-55; discussion 155-6.

Schmidt, M. (2001b). "Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples." Hum Gene Ther **12**(7): 743-9.

Schwarzwaelder, K. (2007). "Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo." J Clin Invest **117**(8): 2241-9.

Sieburg, H. B. (2006). "The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets." Blood **107**(6): 2311-6.

Sieburg, H. B. (2011). "Predicting clonal self-renewal and extinction of hematopoietic stem cells." Proc Natl Acad Sci U S A **108**(11): 4370-5.

Silver, J. and V. Keerikatte (1989). "Novel use of polymerase chain reaction to amplify cellular DNA adjacent to an integrated provirus." J Virol **63**(5): 1924-8.

Smith, L. G. (1991). "Clonal analysis of hematopoietic stem-cell differentiation in vivo." Proc Natl Acad Sci U S A **88**(7): 2788-92.

Stein, S. (2010). "Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease." Nat Med **16**(2): 198-204.

Stewart, M. H. (2010). "Clonal tracking of hESCs reveals differential contribution to functional assays." Nat Methods **7**(11): 917-22.

van Heijst, J. W. (2009). "Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient." Science **325**(5945): 1265-9.

Wang, G. P. (2010). "Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial." Blood **115**(22): 4356-66.

Weber, K. (2011). "RGB marking facilitates multicolor clonal cell tracking." Nat Med **17**(4): 504-9.

Weissman, T. A. (2011). "Generating and imaging multicolor Brainbow mice." Cold Spring Harb Protoc **2011**(7): 763-9.

Wu, A. M. (1968). "Cytological evidence for a relationship between normal hemotopoietic colony-forming cells and cells of the lymphoid system." J Exp Med **127**(3): 455-64.

Zhang, J. (2012). "The genetic basis of early T-cell precursor acute lymphoblastic leukaemia." Nature **481**(7380): 157-63.

Zufferey, R. (1998). "Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery." J Virol **72**(12): 9873-80.