Application of Manufacturing Tools to the DNA Sequencing Process

by
Louis E. Herena

B. S. Chemical Engineering, University of Texas at Austin, 1986

Submitted to the Sloan School of Management and the
Department of Chemical Engineering in partial fulfillment
of the requirements for the degrees of

~~Master of Science in Management~~
and
Master of Science in Chemical Engineering

in conjunction with the
Leaders for Manufacturing Program

at the
Massachusetts Institute of Technology
June 1999

Signature of Author _____
Sloan School of Management
Department of Chemical Engineering

Certified by _____
Professor Charles L. Cooney, Thesis Advisor
Department of Chemical Engineering

Certified by _____
Professor Stephen C. Graves, Thesis Advisor
Sloan School of Management

Accepted by _____
Robert E. Cohen
St. Laurent Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Accepted by _____
Lawrence S. Abeln, Director of the Masters Program
Sloan School of Management

Application of Manufacturing Tools to the DNA Sequencing Process

By
Louis E. Herena

Submitted to the Sloan School of Management and the Department of Chemical Engineering
in partial fulfillment of the requirements for the degrees of

Master of Science in Management and
Masters of Science in Chemical Engineering

Abstract
The state of the art for DNA sequencing has reached the point where it is economically feasible to sequence entire genomes in a matter of a few years. The demand for this data both from public research institutions and private enterprises is tremendous, as evidenced by the entry of several companies in 1998 to challenge the NIH funded Human Genome Project to a race to sequence the Human Genome.

This particular study involves the use of manufacturing strategy and tactics to help a research-based institution such as the Whitehead Institute achieve their production sequencing goals. The findings of this study illustrate the remarkable speed at which new technologies are implemented in the field and subsequent organization and execution challenges that face these high technology centers.

The manufacturing tools applied include constraint management, variation reduction, organizational alignment, quality assurance rationalization and inventory management. In the area of constraint management, a scale-up tool was developed to gain insights of potential problems and opportunities involved in scaling up throughput by a factor of three. Variation reduction was accomplished by the use of better documentation, work standardization, key performance measurement and statistical analysis. The impact of organizational structure was analyzed and cross-training was found to be particularly helpful in advancing knowledge transfer, lowering variability and debottlenecking. Quality assurance was updated for various steps of the process, resulting in potential cost savings. Finally, a model was developed to calculate optimum inventory levels for the core sequencing operation, which will enable more rapid ramp up of new process developments.

The thesis ends with a discussion about the choice of using incremental or radical improvement and concludes that if scale-up data are available, that radical improvement is better for high variability, unstable processes, while incremental improvement is better for low variability, robust processes.

Thesis Advisors:
Professor Charles L. Cooney, Department of Chemical Engineering
Professor Stephen C. Graves, Sloan School of Management

Acknowledgements

The author gratefully acknowledges the support and resources made available to him through the MIT Leaders For Manufacturing program, a partnership between MIT and major U.S. manufacturing companies.

Special thanks to my advisors, Charles L. Cooney and Steven C. Graves for their insights, feedback and pursuit of knowledge.

The author also is indebted to the people of the Whitehead Institute Center for Genome Research for their warm welcome into their world-class research organization, their patient explanations of how their processes work and their openness to different perspectives.

This thesis is especially dedicated to my parents, Ricardo and Nivia, for their love, nurturing and patience through the years. Dedications also go to my brothers Juan, Peter and Robert, my sister Susan and Melissa for their love and support.

Table of Contents

ACKNOWLEDGEMENTS **3**

I. INTRODUCTION **5**

A. BACKGROUND 6
B. STATE OF THE ART DURING THE INTERNSHIP 10
C. CHALLENGE TO THE STATE OF THE ART 11

II. STATEMENT OF THE PROBLEM **14**

III. APPLICATION OF MANUFACTURING TOOLS **17**

A. CONSTRAINT MANAGEMENT IN CORE SEQUENCING 17
*1. **Manufacturing Scalability Model*** *20*
*2. **Results from Manufacturing Scalability Model*** *22*
*3. **Outcomes from Manufacturing Scalability Model*** *23*
B. VARIATION REDUCTION 26
*1. **Core Sequencing Variation Reduction*** *26*
*2. **Library Construction Variation Reduction*** *28*
a) Approach to Variation Reduction and Throughput Improvement 30
b) Initial Assessment of Library Construction Process 31
c) The Plating QC Process 31
d) Use of Controls in Library Construction 32
e) Development of a New Plating QC Predictor 33
f) Correlation of New Empty Vector Predictor for Plating QC 34
g) Reducing Variation of the New Predictor 35
h) Effect of Variation Reduction on New Predictor 37
C. ORGANIZATIONAL STRUCTURE 39
*1. **The Challenge of Functional Organizations*** *39*
*2. **Solving the Functional Organizational Problem*** *40*
*3. **Organization of the Center and Effects*** *40*
*4. **Assessment of Core Sequencing Organization*** *42*
*5. **Outcomes of Core Sequencing Organizational Assessment*** *44*
D. QUALITY ASSURANCE 45
*1. **Quality Assurance in the Core Sequencing Step*** *45*
*2. **Quality Assurance in the Library Construction Step*** *46*
E. INVENTORY MANAGEMENT 51

IV. PROCESS DEVELOPMENT DECISION SUPPORT HYPOTHESIS **54**

A. HYPOTHESIS 54
B. THEORY 54
C. RESULTS 55
D. CONCLUSIONS 55

V. CONCLUSIONS **57**

A. THE USE OF MANUFACTURING MANAGEMENT TOOLS 57
B. FINAL RECOMMENDATIONS 58

VI. APPENDIX **60**

I. Introduction

The Human Genome Project is like no other project ever before initiated by the scientific community due to its scale, timeframe and degree of international collaboration. The Human Genome Project was officially launched in the United States on October 1, 1990, funded by the National Institutes of Health (NIH), the Department of Energy (DOE), the Wellcome Trust and other governments and foundations throughout the world. The project ultimately involves sequencing the human genome as well as several model organisms and developing new technologies to understand gene functionality.

Eager to take advantage of the basic sequencing data provided by the project, most biotechnology, pharmaceutical and research institutions are anxious to speed up the sequencing process as much as possible. Indeed, the impact of private enterprises to more rapidly discover genetic data has influenced the Human Genome Project timeline and approach. The goal at the beginning of 1998 was to have the genome sequenced by 2005. By the end of 1998, partially in response to challenges by the private sector,[1] the project's timetable was accelerated to complete by 2003.[2]

Completion of the project will require running tens of billions base pair analyses. Because of the repetitive, large volume nature of the work, some research organizations call this phase "production sequencing." In order to meet these goals, the researchers involved need to adopt process technologies, innovation and discipline not usually employed in lab settings. By applying some of the manufacturing tools and methods devised in the last two centuries, the goal of sequencing the Human Genome may be achievable within this timeframe.

[1] "Shotgun Sequencing of the Human Genome," Science, Vol. 280 (5 Jun 1998), pp. 1540-1542.
[2] "New Goals for the U.S. Human Genome Project: 1998-2003," Science, Vol. 282 (23 Oct 1998), pp. 682-689

A. Background

The "human genome" is defined as "the full set of genetic instructions that make up a human being."[3] Deoxyribonucleic acid (DNA) is codified using a four-letter system (base pairs A, C, T or G), with three-letter "words" called codons (i.e. AAA, TGA, etc.). Each codon represents either an amino acid or a signal that the protein replication is finished. Since proteins are composed of amino acids, every protein found in nature can be made by following a DNA template. The DNA template for the entire protein is called a gene. It is estimated that there are 80,000 to 120,000 genes in the human genome.

One can analogize the entire three-billion base pair sequence for a particular person to the "source code" of a computer program. Having the "source code" does not convey the functionality of the program unless one understands what the "subroutines" encoded actually do, what makes them execute and what specific outputs they provide. Similarly, knowing the DNA sequence of a person is akin to knowing the entire source code, and the genes are the subroutines of the program. There are two remarkable principles at work here: First, although humans have small differences (1 in 10,000) in their genomes ("source code"), all have the same number of genes, allowing a basis for comparison that can be used to better understand the function of each "subroutine". Second, genes are conserved to some degree in nature. That is, although evolution has forced divergent paths for different organisms, many of the key genes are similar to some degree, allowing study of analogous functions.

Further, the biological principles outlined above can be coupled with the use of technology in order to accelerate the understanding of gene function. Using recombinant technology, it is possible to insert genes or DNA from one organism to another. This technology enables scientists to induce an organism to produce ("express") a protein of interest even if it comes from a foreign gene. Similarly, DNA from a foreign source can

[3] " A Gene Map of the Human Genome: International Group Maps a Fifth of all Genes of the Human Genome", MIT/Whitehead Institute Publication, 1997

be inserted into an organism that can be induced to replicate, thus providing copies of the original DNA. The Human Genome Project is motivated on the belief that having a baseline for comparison between human and model organisms will accelerate and enable gene discovery and understanding of gene functionality.

The NIH organized the Human Genome Project by creating a division called the National Human Genome Research Institute, which coordinates between and provides funding for sequencing centers to decipher certain parts of the genome. The major sequencing centers are the University of Texas Southwestern Medical Center, Baylor College of Medicine, Whitehead Institute/MIT, Stanford University, University of Washington, University of Oklahoma and Washington University. The goals for the period 1993-1998 and the status as of October, 1998 are shown in Table 1:[4]

Table 1: U.S. Human Genome Project Status as of October, 1998.

Area	**Goal (1993-1998)**	**Status (Oct. 1998)**
Genetic Map	2-5 centiMorgan resolution	1 cM map Sept. 1994
Physical Map	30,000 STS's	52,000 STS's mapped
DNA Sequenced	80 million base pairs	291 million base pairs
Sequencing Technology	Radical and incremental improvements	$0.50 per base pair Capillary electrophoresis Microfabrication feasible
Gene identification	Develop technology	30,000 EST's mapped
Model organisms	*E. coli*: complete sequence Yeast: complete sequence *C. elegans*: most sequence *Drosophila*: begin sequence Mouse: 10,000 STS's	Completed Sept. 1997 Completed Apr. 1996 Completed Dec. 1998 9% done 12,000 STS's mapped

The genetic map of the human genome was completed in 1994. The genetic map compares phenotypes, which are the physical attributes that genes convey, for example blue versus brown eyes. This effort produces a genetic (also called linkage) map, used to determine the order and relative distances between genes.

[4] "New Goals for the U.S. Human Genome Project: 1998-2003", Science, Vol. 282 (23 Oct 1998), pp. 682-689.

The mapping of the human genome was done in order to obtain enough information to start the process of sequencing. One main way of physically mapping utilizes sequence-tagged sites (STS's), which are known, short-length sequences (for example AAGCTG) that can be used to roughly find out where a particular piece of DNA belongs.
As shown in Table 1, the project has done very well so far in meeting its sequencing goals, although there was a period where the project was struggling. In fact, as recently as May 1998, there were reports that none of the major sequencing centers had met their two-year sequencing goals.[5] The main reasons for the problems in meeting the goals were the technological and organizational challenges required of step increases in output.

The state of the art in 1993 was such that large scale-up of existing DNA sequencing technologies would be prohibitively expensive. Thus, one of the major goals of the initial part of the project was to help seed advancement of new technologies required to execute the process cost-effectively. There have been many process improvements in the seven years since the Human Genome Project started. Some of the improvements include: higher efficiency recombinant organisms; robotic automation of the preparation procedures; DNA sequence analyzers with higher resolutions and longer read lengths; more robust and standardized data assembly software (informatics); and more refined techniques on preventing and closing gaps.

Indeed, one can characterize the state of DNA sequencing technology to be at the growth part of the S-curve,[6] meaning that there is less effort needed for the same amount of process improvement. It is well known that when technologies reach the growth part of the S-curve, efficiencies and economies of scale become more important than new developments[7]. The evolution of technology is a challenge that all sequencing centers should take seriously, in that attention should be shifted to building economies of scale and productivity.

[5] "DNA Sequencers' Trial by Fire", Science, Vol. 280, (8 May 1998), pp. 814-817.
[6] Foster, R., Innovation, The Attacker's Advantage, (NY: Summit Books, Simon and Schuster, 1986), pp.88-111.
[7] Rebecca Henderson, Notes from Technology Strategy course at MIT (Fall, 1998).

However, as discussed by Foster (op. cited), S-curves usually come in pairs, with the advent of new dominant designs eventually replacing old paradigms. Organizations that over-focus on one S-curve will be at a disadvantage relative to the competition, who may already be one the next generation S-curve. Therefore, while organizations should focus on being productive, they must also be flexible enough to move to the new S-curve as the technology changes.

The NIH recognized that in order to understand gene functionality, they must first find out which proteins are expressed in organisms, as they give great clues as to what parts of the DNA sequence are actually used. These proteins lead to expressed-sequenced tags (EST's), which are sequences that are known to originate from a particular gene because they correspond to proteins that are actually being produced in living organisms.

Finally, as mentioned earlier, the elucidation of DNA sequences of model organisms serves as a platform by which to understand human gene function, due to the similarities in gene function found in nature.

There are additional goals for the Human Genome project for the next five years, including:

- Increase aggregate (all centers) sequencing capability from 90 to 500 Mb/year.
- Decrease cost of finished sequence to $0.25/base.
- Map 100,000 single-nucleotide polymorphisms (SNP's or "snips"), which are differences in sequence from one human to another of one nucleotide base.
- Develop more efficient ways of identifying genes.
- Develop technologies to elucidate gene function.

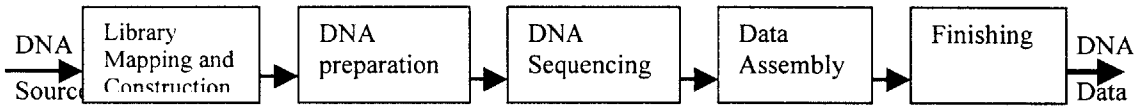
Curiously enough, according to the NIH model, the same organizations that do genetic research such as studying gene functionality will also do production sequencing , which requires entire new competencies focusing on productivity. Therefore, these organizations must build these competencies as well as keep their old ones, becoming more vertically integrated. Such organizations will face many of the same challenges that

pharmaceutical companies do: having to balance two competencies, in their case research and marketing.

This thesis focuses on the goal of high-efficiency (production) DNA sequencing, which was the focus of the internship at the Whitehead Institute Center for Genome Research.

B. State of the Art During the Internship

The NIH, in collaboration with the main sequencing centers, establishes the guidelines for the process of DNA sequencing. The highly accurate approach involved five main steps, shown below:



This process analyzes small pieces of the genome at a time. The source DNA source (also called a BAC clone) is a small, very roughly mapped portion of the entire genome, about 100,000 "base pairs" long. Library Mapping accurately maps the source to a region of the genome by comparing it against known markers. In Library Construction, the DNA source is replicated, purified, sheared into small pieces and presented to the DNA preparation step packaged as a collection of recombinant organisms (a "library" of organisms that, in aggregate, have all of the original DNA source). In the DNA preparation step, the DNA in recombinant form is replicated, purified and molecular "dye" is added to it using PCR technology developed in the 1980's. Once the DNA pieces are prepped, they are sent to sequencing machines that use gel electrophoresis and fluorescence detection to analyze their sequences. The data from the sequencing machines are used to re-assemble the sequence of the entire original piece of DNA. After data assembly, if any gaps remain, that is, if any parts of the original DNA source that for some reason did not sequence, the process is "finished" by applying a variety of special techniques that depend on the nature of the problem. One can look at the finishing step as rework or as a consequence of the limitations of existing technology to produce error-free output. Usually, it is a combination of both, although it is mostly the latter at

Whitehead. The final output is the DNA data or "sequence" of the original DNA source (100,000 base pairs worth). This whole process is repeated tens of thousands of time in order to sequence the entire human genome, which has 3 billion base pairs.

C. Challenge to the State of the Art

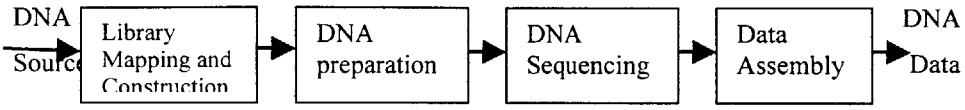
On May 9, 1998, J. Craig Venter, founder of The Institute For Genome Research (Bethesda, MD) and Michael Hunkapiller, president of the Perkin-Elmer's (Norwalk, CT) Applied Biosystems Division announced that they were forming a new company to sequence of the entire human genome in three years at a cost of about $300 million. At the time, the principals of the new venture communicated that the company would try to patent around 100-300 new genes, create a whole-genome database to market to academic researchers and companies on a subscription basis and have a proprietary set of 100,000 single-nucleotide polymorphisms (SNP's), which reveal simple variations in DNA between individuals. This announcement came as a shock to the biomedical research community, which expected to take an additional seven years and expenditures of $1.5 billion to finish the Human Genome Project.

On September 23, 1998, Perkin-Elmer announced the creation of a new business division called Celera, which will trade on the open market as a targeted stock: "Its mission is to become the definitive source for genomic and related biomedical information. Celera's plans include: 1) sequencing (draft) of the complete human genome during the next three years; 2) discovering new genes and regulatory sequences that could comprise new drug targets; and 3) elucidating human genetic variation and its association with disease susceptibility and treatment response. Celera plans to create commercial value through the license or sale of databases, knowledge-bases, proprietary drug targets and genetic markers, and related partnership services."[8] Trading Celera as a targeted stock, rather than an independent company, presumes that Perkin-Elmer wants to keep close managerial control of and add synergy to the new enterprise. Additionally, by keeping

[8] PERKIN-ELMER ANNOUNCES PROPOSED DISTRIBUTION OF CELERA GENOMICS TARGETED STOCK", Perkin-Elmer Corporate Public Announcement, NORWALK, CT, September 23, 1998.

close ties with Celera, Perkin-Elmer will increase their absorptive and commercialization capacity into instrument systems and reagents, their core businesses. Perkin-Elmer plans to internally subsidize the new stock issue, signaling high confidence in the venture.

Celera proposes to eliminate the labor-intensive steps of library mapping and eliminate finishing altogether:



This process sequences the entire genome in one shot. The DNA source is the 3 billion base pairs that make up the entire human genome. The middle of the process is similar to the NIH approach, but with much less generation of data per DNA source. In Library Construction, the DNA is replicated, purified, sheared into small pieces and presented to the DNA preparation step packaged as collection of recombinant. In the DNA preparation step, the DNA in recombinant form is replicated, purified and molecular "dye" added to it using PCR technology. Once the DNA pieces are prepped, they are sent to sequencing machines to analyze their sequences. The data from the sequencing machines are used to re-assemble the sequence of the entire original piece of DNA. The "gaps" are not finished and the final output is the entire DNA genome data.

In addition to eliminating the portions of the NIH process, Celera plans on using Perkin-Elmer's new capillary electrophoresis machines for the DNA sequencing. However, versions of these capillary machines (made from Perkin-Elmer's competitors) are already available to the NIH centers.[9]

The impact of the Celera challenge was to change the NIH approach to obtain a rough version of the genome, with continuing refinement in resolution to come at a later date. It

[9] "Sequencing the Genome, Fast", Science, Vol. 283, (19 March 1999), pp. 1867-1868.

is in this challenging environment that the Whitehead Institute entered their third year of DNA sequencing operations.

II. Statement of the Problem

The problem facing the Whitehead Institute's Center for Genome Research was to scale up their operations while keeping high quality output and meeting cost targets outlined by their research grants. Additionally, they had various development projects in their pipeline aimed at minimizing the labor costs, lowering reagent costs and increasing efficiency. Their goal for June 1998-June 1999 was to sequence 16 million base pairs at a cost of $0.50 per finished base pair. The purpose of the internship was to provide manufacturing perspective and knowledge to their research-oriented culture by helping them formulate and execute various manufacturing strategies. This is relevant as they enter the production sequencing phase of the project.

This thesis will trace various manufacturing strategies implemented at the Whitehead Institute, where due to the recent scale-up of the Human Genome Project, parts of the Center for Genome Sequencing was transitioning from a research to a factory culture at an accelerating rate.

In particular, this thesis will examine two areas in detail:

1. **Application of manufacturing tools** – what worked, what did not and why. The choice of manufacturing tools was based on discussions and agreements with local management.
    a. Constraint management:
        - Core Sequencing Manufacturing Scalability Model
    b. Variation Reduction:
        - Core Sequencing
        - Library Construction
    c. Organizational structure:
        - Analysis and recommendations for Core Sequencing
    d. Quality Assurance:
        - QA in Core Sequencing

- QA in Library Construction

e. Inventory Management:

- Effect on inventory on throughput and development speed.

The approach to address the manufacturing concerns was to have an initial kick-off meeting with the Center's top management in order to discuss the burning issues. At the conclusion of the first meetings, we decided that objectives for the project were first to build a model of the core sequencing operations, and then to use the model to bring a manufacturing perspective by executing it in a variety of projects to increase output and decrease cost.

The approach was to apply manufacturing tools in operational areas deemed to be ready for production sequencing. Although the overall system was constrained in the finishing operations, the main emphasis was to elevate the constraints of the core sequencing operations, followed by the library construction area. The main point of the thesis is to evaluate the impact of various manufacturing tools on the output and efficiency of the operation. Since there were a variety of improvement efforts going in parallel, it is difficult to separate the effects that these tools have. Therefore, the evaluations of how well the tools worked will be more qualitative in nature.

2. Process development decision support hypothesis

How does an operations manager choose between an incremental or radical improvement effort? In addition to throughput and cost considerations, operations leaders must sometimes choose between committing resources for radical improvement efforts or incrementally improving the process. Although constraint theory helps pinpoint where to apply resources, it does not address the next decision: to try radical or incremental improvement.

In a process where many bottlenecks and scale-up considerations constrain the process, the operational manager must assign limited development resources so that they deliver

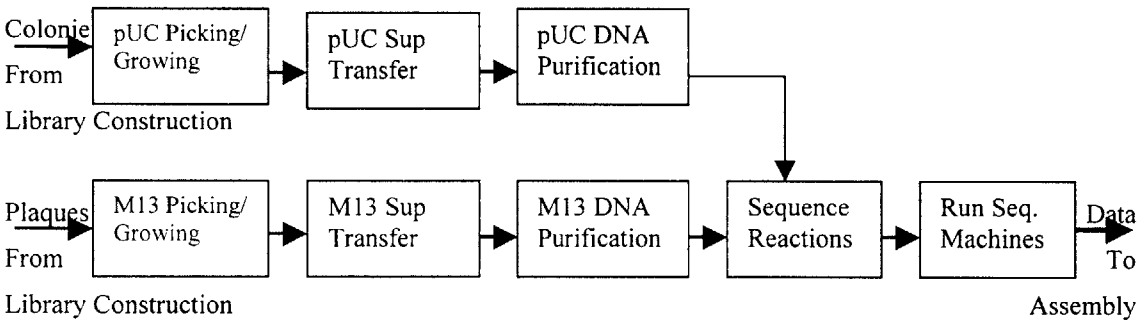
throughput improvements at the appropriate time. The choice of addressing improvements as incremental or radical is a matter of technical and organizational judgement. Nevertheless, a framework with which to think about this problem could help to make better decisions. The hypothesis assumes that scale-up data are available for a new process, which would provide a radical improvement in terms of productivity. If the process has high variation, making it more difficult to make incremental improvements, the decision should be to try the radical improvement. Conversely, if the process has low variability, the decision should be to do incremental improvement, as it is less disruptive and more economical.

An attempt will be made to quantify the variability of various processes, classify the improvement efforts as either radical or incremental, evaluate the success of the efforts and attempt to prove or disprove the hypothesis based on the data from this internship.

III. Application of Manufacturing Tools

A. Constraint Management in Core Sequencing

We first concentrated our efforts in the "shotgun" or core sequencing steps, which includes DNA preparation and sequencing. This process is by far the most automated section of the facility, with much of the laborious tasks performed by programmable robotic arms and automatic dispensers. The core sequencing operation is summarized by the following:



M13 Picking/Growing

The raw material for this process comes in the form of plaques from Library Construction. Plaques are made by spreading individual M13 phages (each phage has some human DNA inserted into it) onto a plate coated with *E.coli* (called "lawn cells") and nutrients. The individual phages (M13's) infected and burst neighboring lawn cells, creating holes in the lawn of cells. Each plate contains from 50 to 500 plaques. Each plaque corresponds to an original M13 that had a small piece of human DNA (about 2000-2500 base pairs long) inserted into it. The plaque is "picked" by touching it with a toothpick or similar instrument, which transfers infected cells and phages onto its surface. The instrument is then dipped into a growth media with *E.coli*, where some of the phages are shed from the surface and allowed to replicate for about 16 hours at 37 degrees Celsius. For a typical DNA fragment ("project") of 100,000 base pairs, 1200 of these plaques are analyzed.

pUC Picking/Growing

The raw material for this process comes in the form of colonies from Library Construction. Colonies are made by spreading individual *E. coli* cells infected with plasmids (each plasmid has some human DNA inserted into it) onto a plate coated with nutrients. The individual cells replicate both themselves and the plasmids they are infected with, creating colonies (clones) of the original plasmid. Each plate contains from 50 to 500 colonies. Each colony corresponds to an original plasmid that had a small piece of human DNA (about 2000-2500 base pairs long) inserted into it. Each colony is "picked" by touching it with a toothpick or similar instrument, which transfers infected cells onto its surface. The instrument is then dipped into a growth media with *E.coli*, where some of the cells are shed from the surface and allowed to replicate for about 16 hours at 37 degrees Celsius. Towards the middle of the internship, for a typical DNA fragment (also called a "project") of 100,000 base pairs, 1200 of these colonies are analyzed. At the beginning of the internship most of the shotgun operation consisted of M13's, with only about 10% of a project analyzed using pUC's, using about 2160 M13's and 240 pUC's per 100k project.

M13 Supernatant Transfer

The cells in the growth media encourage the replication of the M13, which infects the surrounding cells and eventually bursts them. The M13's end up in the supernatant phase and they are isolated from the *E.coli* cells by "spinning" the growth plate down in a centrifuge. 100 μL of the supernatant is added to 10 μL of 20% SDS solution (surfactant), providing a stable solution for freezer storage. At this point, the samples are in 96-well microtiter plates, where 16 well out of each plate (96 wells) are sampled and tested for adequate DNA content for further processing. The test consists of running the samples on gel electrophoresis, where the criteria to pass is to have less than 4 out of 16 with low DNA. If a plate does not pass QC, it is discarded and replaced with new supernatant.

pUC Supernatant Transfer

The cells in the growth media encourage the replication of the pUC-infected *E.coli*. At the end of the growth phase, the growth plate is spun down with a centrifuge and the supernatant discarded. The resulting "pellets" are placed in 96-well microtiter plates ready for purification.

M13 DNA Purification

At this point, the DNA is sent to the purification step, where the purpose is to isolate the DNA from the rest of the growth material. This is done using a technique called SPRI (solid-phase reversible immobilization), where under the presence of certain reagents, DNA will bind to carboxyl-coated magnetic particles. When this happens, the original solution can be discarded and the DNA washed with solvents. After the wash step, the DNA is released from the magnetic particles and stored in an aqueous solution.

pUC DNA Purification

The purpose of this step is to separate the pUC DNA from the *E. coli* DNA. This is done using a proprietary technique developed at Whitehead. The end result is similar to the M13 purification, with only the uncircularized plasmid DNA remaining, ready to be dye-replicated using PCR.

Sequence Reactions

The purified DNA is now ready to be processed for sequencing reactions. One of the methods used is the Sanger technique,[10] where the DNA is replicated using PCR technology in the presence of dye oligonucleotide primers. This causes a nested set of end-dyed DNA to be produced. Since there are four bases in DNA (A, C, T and G), each base has a different dye. Each microtiter plate (with 96 wells) is split into a 384 well plate, the reactions performed and the 384 well plate "repooled"' back into a 96 well plate. At the end of this step, the DNA sample is ready to be run through the sequencing machines.

[10] An Introduction to Genetic Analysis, Anthony J. F. Griffiths et al, pp. 446-447, 6th edition , 1996 W.H. Freeman and Company

Sequencing Machines

The dyed-DNA sample is loaded into a "sequencing machine," made by Perkin-Elmer, model ABI-377. The sequencing machine uses a combination of gel electrophoresis and a fluorescence detector. Electrophoresis separates molecules in an electric field by virtue of their size, shape and charge. If a mixture of DNA molecules is placed in a gel matrix with an electric field applied to it, the molecules will move through the gel matrix at speeds dependent on their size. The smaller molecules will move (elute) faster and so the smallest string of the nested DNA set elutes first. There is a laser and a detector that measures the fluorescence of the sample as it elutes. This provides an output similar to a gas chromatograph, which is then interpreted using software. This machine is capable of running 48 to 96 wells (samples) at a time in a period of about 12 hours (including run and loading time), giving average read-length of about 800 base pairs or 1600 base pairs per lane per day. Alternatively, the machine can be run in an 8-hour cycle time, but the read-length drops down to 600, or 1800 base pairs of output per day per lane. which makes for more difficult data assembly and processing. Although from a strict "output" view, it would seem that it is better to run the machines three times per day, studies at Whitehead showed much better assembly data (less "defects") from the longer read-lengths, partially due to the long repeats region that are sometimes encountered. Thus, Whitehead ran the 12-hour cycle. All of the data from these machines is sent to a central data base, which collects data from all 2400 samples of the project and comes out with an estimate of the sequence of the original DNA fragment (100,000 base pairs).

1. Manufacturing Scalability Model

One of the most important questions at the beginning of the internship was regarding the scalability of the existing process. The operations management felt that although they had a great deal of automation in place (accounting for about 25% of all unit operations) and more automation under development, they wanted to know the effects of scaling up their current model. The main philosophy was to keep the number of personnel low and utilize automation to increase capacity. The sequencing machines were immediately

identified and validated to be the bottleneck step, running 5 days a week, 24 hours per day (5x24). The other steps were run in one shift and so had plenty of spare capacity for the short-term needs.

However, there were many near-term changes in the works. The bottleneck step was undergoing changes that would significantly affect the capacity of the entire system. First, the number of wells per machine was increasing from 64 to 96 and the number of machines increased from 19 to 40. This would effectively increase the bottleneck capacity by a factor of three in the next six months. In the meantime, the ratio of M13 to pUC's (plasmids), previously being 10:1, increased to about a 1:1 ratio, essentially requiring a new automated purification process. In addition, the core sequencing step picked up a lot of the finishing operation capacity, due to its economies of scale, adding a degree of complication to coordinating and prioritizing daily operations. Lastly, the operation had to have the capability of quickly changing reagent mixes, which due to their high costs, were continuously being optimized. The model in its final form is shown in Exhibits 1, 2 and 3, which are linked spreadsheets.

Exhibit 1 breaks down every manual or automated step and shows the setup and process time for each, and estimates the labor required to perform each batch (which depends on the batch size). Exhibit 2 then uses this information to summarize the labor requirements for each major area of core sequencing.

Exhibit 2 is the master spreadsheet, which takes data from the designed bottleneck of the plant, the ABI sequencing machines. The output of the ABI's is determined by the number of machines assigned to core sequencing, the number of lanes run per machine and the gel cycle time. The number of machines dedicated to core sequencing was set by the total number of machines minus the number of machines down for maintenance at any given time, minus the machines needed to run library QC, finishing and development samples. Running at 96 lanes per machine took some time to implement because the gel geometry was fixed, which decreased the lane clearance, requiring optimization of the upstream and downstream processes. The gel cycle time was generally fixed at 12 hours,

although it was set up as a variable for further studies. Once the number of plates that could be run per day was set, the batch size for each operation was set, which then gave the number of batches per day required for each operation.

Exhibit 3 takes the information from Exhibit 2 and automatically generates a schedule of events for each batch (called a "cycle" on the spreadsheet), estimating the amount of time required to perform each major core sequencing operation. Exhibit 3 was particularly useful in evaluating alternatives for one-shift operation by quickly pointing out when the number of batches and associated cycle times exceeded an eight-hour day.
Exhibit 2 also linked the "coverage pattern", which is the number of pUC's (called DS for double-stranded DNA) and M13's (called SS for single stranded DNA) per project and the type of dye chemistry used in each (FP=forward primer, RP=reverse primer, FT=forward terminator, RT=reverse terminator). The coverage pattern was determined by Whitehead to have a radical effect on the number of gaps per project at assembly. However, changing the coverage pattern implied changes in flows through core sequencing, which then required operational adjustments.
The idea behind the model was to identify problems that may come up due to scale-up and process changes, and to assess labor productivity.

2. Results from Manufacturing Scalability Model

The results from the model showed some significant future problems associated with scaling the operation from 19 to 40 ABI machines, coupled with an increase in number of lanes from 64 to 96 wells/machine (three-times increase in scale). The following findings summarize the results:

- Low utilization of personnel for the sequencing machines. At the time the model was developed, the Center had gone from three eight-hour cycles per machine to two twelve-hour cycles per machine with the same number of machines (19). Therefore, the amount of labor required to operate the machines dropped by 33%. Further, the labor efficiency with three shifts was about 70%. The model correctly predicted that the existing crew could run twice the number of machines. However, this would

require a turnaround time of two hours, which would mean that more people would be required during the critical turnaround time (7 a.m. and 7 p.m.). This pointed out that the shifts would have to be re-balanced or cross-trained.

- The picking operation was confirmed to be very labor intensive, utilizing 12% of the labor costs. Although the picking operation had an existing automated picker, it was not used due to technical problems. The labor utilization was already high in this area, showing that the step increase in production would tax the existing crew. This emphasized to the need to either get the existing automated picker on-line or scale up the number of people doing this operation.
- The quality control operations took up a significant amount of labor (8% of the total), emphasizing the need to rationalize it. The scheduling spreadsheet showed that, with the existing rate of sampling, an additional partial shift would have to be added.
- The amount of work-in-process (WIP) inventory was significant, with over six weeks in process compared to a cycle time of three days required with no WIP. This had the effect of making it difficult to quickly see the effect of process changes, forcing "short-circuiting" to get development runs through. Again, this pointed to the need for evaluating and establishing a target WIP inventory.
- In addition, the model predicted that an additional shift would have to be added (assuming no automation added to compensate) to the purification step and the sequencing reaction step due to the increased number of batches to be run per day.

3. Outcomes from Manufacturing Scalability Model

The management of the Genome Center agreed with the insights as presented and agreed to address the potential problems in the following way:

- Increase the cross-training amongst the sequencing machines operators, in order to address the turnaround time problem and allow for better knowledge transfer that would eventually lead to lower variance between machines and allow fairer allocation of work. The allocation problem does not show up in the model, but the way the sequencing machines were staffed, certain people did the gel prep work, while others did the machine loading work, leading to some concern among the crew about having

to do repetitive, narrowly defined tasks. There was a fair amount of interdependence between tasks and it was difficult to account for the reason for gel problems due to the separation of tasks. It was believed that if everyone had a set number of machines assigned to them, they could do the prep work and loading and therefore have more control over the process.

- Have the automated picker "Flexys" system sent back to vendor for repairs. After the picker returned, it was found to be helpful, but not as efficient from a yield perspective as doing the picking by hand, and it still suffered from technical glitches. At the time, library construction (the source of raw material to the core sequencing operation) was barely keeping up with production, and had become the new bottleneck of the operation. Therefore, it was deemed more important to have high library yields and this operation was kept as manual.
- The quality control issues warranted further investigation, the results of which are shown in subsequent chapters. The final outcome was a reduction by 50% of the labor needed for quality control.
- The amount of WIP inventory was a controversial issue, in that the shift supervisor felt a need to keep buffers between steps to minimize the possibility of downtime. The WIP inventory consisted of micro-titer plates stored in refrigerators after each step of the process. There was about two weeks worth of production stored after supernatant transfer, another two weeks after purification and two weeks of sequenced DNA storage. Due to the small size of the samples, it was not perceived to be a large problem, but as is well-known in operations management, served to hide a variety of problems including machine unreliability and absenteeism. This problem was especially exacerbated by the fact that besides the shift supervisor, there was only one person trained to run the purification automation. The sequencing automation had similar staffing problems, with only one person who knew how to run it. The operations manager agreed to the concept of having an appropriate amount of inventory. This issue was studied further and the results shown in subsequent chapters. Due to the ramp-up of core sequencing output coupled with low library construction output, core sequencing ended up running with low inventory de facto.

- Finally, the management decided to speed up the automation in the purification and sequencing reaction steps in order to have a "one-shift" operation. These changes were implemented over a period of time and took up a considerable amount of development time. The advantage of doing this was that it kept the development personnel, who had to respond to production problems, from having to split up their shifts to provide more than one-shift coverage.

B. Variation Reduction

1. Core Sequencing Variation Reduction

Although the benefits of variation reduction were known to the Center, the variability in their processes was not measured on a daily basis. Rather, the Center relied on large excursions from the mean to react to problems. One of the reasons is that the processes were almost never locked down and it was recognized that some of the processes were not in statistical control. Variation reduction in core sequencing was considered important because it provided a way to find throughput and quality problems. As discussed earlier, the significant inventory levels created long lag times (1 day to six weeks) between the source of variation and the final results. In addition, there was inconsistent documentation using lab notebooks, making it difficult to "data-mine" at the lowest operational levels. In an effort to better trace sources of variability, the following plan was already being implemented by the Center:

- Structure documentation similar to that used in the pharmaceutical industry – SOP's (protocols), batch or shift records and change control forms.
- Track machine reliability by manual documentation of failures and uptimes.
- Track key, relevant quality and output statistics for each project.
- Assign development efforts to address major sources of variation.

Outcomes for Core Sequencing Variation Reduction

- Documentation was improved over the existing lab notebook method, batch tracking data sheets similar to current Federal Drug Administration (FDA) good manufacturing guidelines (cGMP's) were used. Protocols were kept more up to date. Change control remained less formal, due to the flexibility requirements of the process.

- Machine reliability tracking was more formal and closer management review than previous, with a feed into the development group for fixing machine problems (most of the machines were specified and installed by the development group).

- Key statistics were tracked on a daily basis, as shown in Exhibit 4, which was a network-accessible file. Exhibit 4 was the main core sequencing tracking sheet, kept updated by the relevant production leads. The tracking sheet was used to coordinate amongst the various groups and provided management with a one-page summary of project status. Exhibit 4 also summarized the quality of assembled data from each project by the following metrics:
  - Overall pass rates – percentage of "reads" of a project that were of acceptable library and sequencing quality.
  - Sequencing pass rates – percentage of "reads" of a project that were of acceptable sequencing quality, implying that the "core sequencing" process described above worked successfully.
  - Library pass rates – percentage of "reads" of a project that had adequate DNA inserted into the sample.
  - Average read length – this gives an indication of how long a string of DNA was read on average for a given project. Generally, the better the quality of the data and the higher the pass rates, the longer the read length.
  - Gap data – after a project is assembled (all the reads done and compiled to get an estimate of the DNA sequence), there were gaps that needed to be resolved. These gaps required manual intervention by the "finishing" group, who had to find the appropriate strategy for resolving the problem and then sent the orders to the lab to process the samples. This generally added a lot of time to the project cycle and it was very desirable to minimize this. Generally, the better the coverage (successful reads per base of DNA fragment), the lower the gaps. During the period of the internship, the Center discovered that the right "coverage pattern" of pUC's, M13's and dye chemistry provided the minimum number of gaps per project.

Exhibit 4 also shows the segregation between groups of projects as coverage patterns or new technologies were introduced into production. The above metrics were continually monitored to measure the impact of major process changes.
The final data for every project were available in a separate web-based system in details ranging from aggregate project statistics down to the exact output of each sequencing machine for every sample.

Although the in-process data remained accessible only by the manual record keeping, there were plans to have these data available for the next-generation automation platform that the development team was working on.

- The variability of the above quality parameters was never formally measured, although this was done on an individual basis for evaluating development projects. The following table summarizes the variability of each quality statistic using data from Exhibit 4. The variability of each statistic is measured using the Coefficient of Variation (Cv), defined as the sample standard deviation divided by the sample mean.

Table 2 – Summary of Key Quality Statistics For Core Sequencing

Time period	% Ovl. Pass		% Seq. Pass		% Lib. Pass		Read Lgth.		No. Gaps	
	Mean	Cv	Mean	Cv	Mean	Cv	Mean	Cv	Mean	Cv
1/98 – 30 projects	74	0.114	80	0.084	93	0.032	547	0.062	9	0.675
2/98-3/98 –20 projects	72	0.105	78	0.057	93	0.047	657	0.062	8	0.689
4/98 –14 projects	72	0.047	80	0.063	93	0.088	692	0.053	7	0.554
5/98 – 12 projects	87	0.070	93	0.046	94	0.022	777	0.039	N/A	

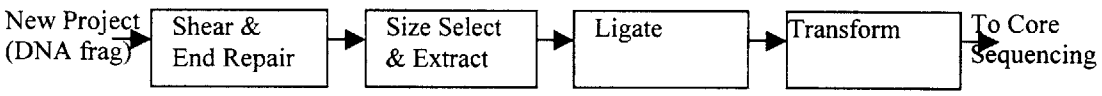
As can be seen in the table above, there was a large increase in sequencing pass rates in May, mostly due to addressing a recurring automation problem associated with the purification system, which had the effect of decreasing variability as well. Although the library pass rates had a steady mean due to the selection process (discussed in the Quality Assurance chapter) imposed on the system, its coefficient of variability changed significantly from month-to-month, by a factor of four from April to May.
The read length increase and variability decrease was due, respectively, to changing over to twelve hour cycle times on the sequencing machines and an internal effort to improve gel quality. The number of gaps per project dropped significantly during this time period and continued to drop throughout the year due to the Center's focus on optimizing the coverage pattern of projects.

2. Library Construction Variation Reduction

Library Construction is the step upstream to Core Sequencing, where the plaques and colonies that contain the DNA fragments of interest are generated. The Center's

management was concerned that this part of the production step would not be able to keep up with Core Sequencing once they ramped up to full production. Core Sequencing was scheduled to be capable of 50 projects per month by the end of the year while Library Construction had averaged about 12 per month from January to May 1998. The concern was not only with scaling the existing operation, but also improving its reliability.

The following flowchart shows the library construction process:



A project is a collection identical DNA fragments, originated and purified from bacteria artificial chromosomes (BAC's) which are clones containing a piece of DNA approximately 100,000 base pairs long. The goal is to break up the large fragments into random, small pieces approximately 2000 base pairs long and package them up (ligate) with a vector such as M13 or pUC. By ligating the fragments to a vector and infecting host cells, known as competent cells, the plaques or colonies formed can be processed in Core Sequencing.

Shearing and End Repair

Shearing of the large fragments is accomplished with an ultrasonic tip inserted into the DNA solution for about 10 seconds at a set power setting. Shearing breaks up the DNA into random sized pieces, but since DNA is double-stranded, a lot of the DNA end up with ends that are single stranded. Single stranded DNA will not ligate and must be "repaired", by adding mung-bean nuclease (MBN). MBN attacks the single-stranded DNA ends by cutting them back until a double strand is found. The reaction is controlled by specified time and mole ratios.

Size Select and Extract

The sheared and end-repaired sample is placed in an agarose gel matrix, using electrophoresis to size select. Voltage applied across the gel box starts the migration of the sample towards one end, with the smaller DNA molecules travelling faster. A marker

of DNA of known length is run at the same time. The end result is a gel streak that can be cut out to select DNA sizes of a certain range (1.6 to 2.4 KB). The first "cut" is then re-processed using the same procedure to get a second cut with a narrower range (1.8 to 2.2 KB). The second cut is then extracted using solvents to clear out any gel remnants and end up with a pure DNA sample, which is tested one more time before ligation.

Ligate

The next part of the process is ligation, where the DNA is fused or packaged with a vector (M13 or plasmid) under certain conditions. The efficiency of this step is a function of many variables including reactant and enzyme mole ratios, ligase activity and time. The reactants are the DNA fragments and the vector. The enzyme is ligase with a buffer to provide ATP for the reaction. The reaction is carried out at 16C overnight.

Tranform

Tranformation is the process by which the ligated vector (vector with DNA insert) infects a host organism in order to replicate it to have enough pickable cells. This step also separates the ligated vectors from each other and unligated vectors. If the ligation and transformation is successful, there will be enough infected hosts to provide an adequate "coverage" of the project (at least 2400 colonies or plaques for a 100 KB project). The ligation/transformation is quality controlled by performing a "Plating QC," where about 10% of the ligated material is transformed. If enough pickable colonies or plaques form (at least 240 for a 100 KB project), the project is deemed adequate as far as coverage and the test transformation is sent on to production for Sequencing QC (discussed in the Quality Assurance chapter). If the test transformation passes the Sequencing QC, the entire project is transformed (also called "plated out", since the transformation process is done on agar plates).

a) Approach to Variation Reduction and Throughput Improvement

In July of 1998, we met to discuss an approach and the following points and questions were posed:

- The process seems go through periods of spectacular success and failure. If the process could be made to run like it does during the successful periods, there would be plenty of capacity to provide for production. Since the process was purely biological (reagents and raw materials), there were many sources of variation. What are the sources of variation? How can they be reduced?
- The lead lab technician for this step was leaving within a month, what was the best strategy going forward?
- Are policies and procedures appropriate? Are they being followed?

b) Initial Assessment of Library Construction Process

One of the things that stood out in this process was the tremendous amount of rework that occurred. The rework was routed at the two main QC points: Plating QC and Sequencing QC. Plating QC, part of the Transformation Step, was performed by the Library Construction team and since 50% of the projects failed at that point, compared to 25% failure rates at the Sequencing QC step, it seemed to be a good potential starting point.

Library Construction was functionally organized, with one lead lab tech who evaluated the Plating QC results, collected Sequence QC data, ordered and tested raw material, assigned daily tasks and filled in when needed; one lab tech who did the first three steps; and two lab technicians that performed the transformations. Since the lead lab tech was leaving soon, it was considered important to understand her decision process for Plating QC. After discussions with the lead tech, it was clear that she used her experience and tacit decision rules to deem whether a project would pass Plating QC. We decided to develop a more robust model that took into account more of the available lab data in order to make more consistent decisions.

c) The Plating QC Process

The following outlines the existing Plating QC process:

- A test transformation was done, giving a number of white plaques and blue plaques. The white plaques presumably were vector with a DNA fragment and the blue plaques were vectors with no fragment (also called empty vector).
- The white and blue plaques were then counted and as long as the white to blue ratio was deemed high enough (5 to 10, depending on how other transformations were working that week and production urgency) and there were enough whites to cover the project, the project was approved. Historically about 50% of projects passed test transformation.
- If a project did not pass, it was sent back to be retransformed, religated or completely reprocessed (again, depending on the conditions at the time).

We wanted to find a better and more quantitative tool to use for Plating QC because we felt that if the QC could be made more accurate and less variable, there would be less Type 1 errors (rejecting when the sample was acceptable) and thus higher throughputs through the system. Further, we wanted to reduce the variability in output due to changing QC test parameters.

d) Use of Controls in Library Construction

For every project that was transformed, there were three controls that were supposed to be run:

Vector alone – to check that the vector, lawn and competent cells were not contaminated.

Vector + Ligase – vector was treated so it would not ligate onto itself – this would check that this was true, gave a baseline of blues and whites to which the main sample could be compared

Vector + Ligase + Calf Thymus (CT) DNA – a known DNA fragment from Calf Thymus was ligated to the vector and it was expected to give a large number of whites, this checked the ligase activity.

The sample itself was run with Vector + Ligase + Project DNA.

We found that some of the controls were not run and the ones that were run, not well documented. The missing controls made it difficult to pinpoint Library Construction

problems when they occurred. The Library Construction management re-emphasized to the laboratory technicians the importance of controls, which alleviated the problem. The ideal Plating QC procedure would incorporate some or all of these controls in the decision process.

e) Development of a New Plating QC Predictor

One of the purposes for Plating QC is to be able to predict the percentage of white plaques that would end up with no DNA fragment. These plaques were also called "empty vector" or "Seq Vector" and it was desired to have less than about 8% of these per project. The other purpose for Plating QC is to estimate the yield of plaques from a particular project in order to verify there will be enough plaques generated to "cover" the project (recall that about 1200 plaques are required per 100 KB project). Since, except for variability in Plating QC, it was relatively easy to determine the yield of the transformation, we concentrated on finding a better predictor.

We first classified the white plaques into two categories:

(1) $W = Wf + Wv$, where

W= total number of white plaques from sample,

Wf = number of white plaques with a DNA fragment and

Wv= number of white plaques that are empty vector

(2) $Wv/W * 100\%$ = Percentage of Empty Vector

However, since it is not possible to tell which of the white plaques are empty vector, we used the Vector+ Ligase control to estimate it.

Let B= number of blues plaques from sample,

Wc= number of white plaques from control and

Bc = number of blue plaques from control

Now we know that if the transformation had been equally efficient for both the sample and the control, B would equal Bc. However, this is not usually the case, but we can use the ratio of B to Bc to get an idea of the relative efficiencies. Similarly, Wv would equal Wc if the transformation efficiencies were the same. However, their relative efficiencies can help us establish the following relationship:

(3) B/Bc=Wv/Wc= Es/Ec, where

(4) Es = sample transformation efficiency and

(5) Ec = control transformation efficiency.

Solving for Wv ,

(6) Wv=Wc(B/Bc) and substituting into equation (2) gives us:

(7) % Empty Vector=(B/W)(Wc/Bc) * 100%

Thus, we had a quantitative predictor that we could use for Plating QC and that utilized a closely associated control.

f) Correlation of New Empty Vector Predictor for Plating QC

We attempted to apply the data we had available to run a linear regression of equation (7). The two independent variables were B/W and Wc/Bc, both obtained from plating QC data from all 39 projects from June through August for which we had data. The dependent variable, % Seq Vector, was obtained from Sequencing QC data. The regression is plotted in Exhibit 5, showing a good correlation to % Seq Vector (SV-Empty Vector). Theoretically, the coefficient of the regression should have been 1, but instead we obtained a value of 0.5, with a 95% confidence of 0.4 to 0.6. Although we did not find a reason for the difference, we found that the equation correlated better to the dependent variable being % Seq Vector + % Small Ins. % Small Ins is the percentage of very small DNA inserted into the vector. The new regression showed a better correlation and a more reasonable value for the coefficient (1.2), with the theoretical value of 1

falling within the 95% probability limits. The ability to predict %SI along with %SV was considered to be an advantage and although our theory did not predict this would happen, we decided to use it as an empirical tool anyway, since it was considered to be better than the current method.

g) Reducing Variation of the New Predictor

The next step was to try to reduce the variation of the predictor itself. We made the hypothesis that B/W and Wc/Bc were independent, ran a linear regression between the two variables and found no correlation between them. Since our predictor is a product of two independent variables, the coefficient of variability of our predictor could be estimated using a method discussed by Himmelblau[11]:
For a general equation involving a product of many variables,

$Y= a*X1*X2...Xn,$

Where, "a" is a constant, X1 is factor variable1, X2 is factor variable 2, and Xn is factor variable n, Himmelblau shows that

$(Cv,Y)^2 = (Cv,X1)^2 + (Cv,X2)^2 +...+ (Cv,Xn)^2$, where

{Cv,Y} is the coefficient of variation (Cv) for the variable Y, {Cv,X1} is the Cv of X1, {Cv,X2} is the Cv of X2, and {Cv,Xn} is the Cv of Xn.

Applying the above equation to our predictor gives us:

$(Cv,SV+SI)^2 = (Cv,B/W)^2 +(Cv,Wc/Bc)^2,$

We noticed that since the control usually resulted in a low number of blues (Bc) and whites (Wc), with values ranging from 1 to 10 and 5 to 50 respectively, the low counts could be contributing a high proportion of the overall variability. We further assumed

that Wc and Bc are binomially distributed, that is, for any given sample of the vector, there is a constant fraction of blue (p) or white plaques (1-p) that would appear. Assuming the controls Wc and Bc are binomial, a way to reduce the Cv,Wc/Bc term is to take a larger sample. Recall that:

Cv= Sx / X , where Sx is the standard deviation and X is the mean of the distribution.

And for a binomial function,

X = np and Sx^2=np(1-p)

where n is the number of outcomes and p is the probability of the outcome.

Therefore ,

(Cv)^2= (1-p)/np , and as the number of outcomes increases, the square of the coefficient of variation decreases and therefore so does the contribution of variability from the control.

Reducing the variability from the sample (B/W) would be more difficult, as there were many more potential contributors to it. After some discussion with management, we decided to try the following plan to reduce variability:

- Update protocols to reflect existing practice.
- Perform more formal cross-training.
- Keep more accurate batch records.
- Gather process data to find correlation with failures.
- Stricter adherence to protocols and controls, be more consistent with process times and batch sizes.

The following table summarizes the variability found in Library Construction over the time these changes were implemented:

[11] "Process Analysis and Simulation – Stochastic Systems" D. M. Himmelblau, University of Texas at Austin, 1969 pp. 38-39.

Table 3 – Summary of Variation In Library Construction

Time period	B/W Mean	B/W Cv	Wc/Bc Mean	Wc/Bc Cv
June-Aug, 1998 (167 samples)	0.71	1.79	0.61	0.94
September, 1998 (73 samples)	0.43	1.38	0.28	0.56
October, 1998 (96 samples)	0.39	1.00	0.39	0.74
November, 1998 (63 samples)	0.48	0.82	0.35	0.36
November, 1998 (11 HS samples)	0.08	0.88	N/A	N/A

The mean B/W ratio for the months of September through November did not change. However, the coefficient of variation decreased from approximately 1.4 to 0.8 in the same time period, indicating that there was some reduction in variation. The mean Wc/Bc ratio remained relatively constant from September to November. The Wc/Bc coefficient of variation did not show any clear trends indicating either an increase or decrease in variability. In November, a new procedure (Hydrashearing- HS) for processing the DNA was implemented on a trial basis and showed great promise, as shown by the dramatic decrease in the mean value. A decrease in the B/W ratio is desirable, since it indicates more whites per unit blue. Although the new procedure had about the same coefficient of variation as the old procedure, its standard deviation was much lower (due to reduction of its mean value).

h) Effect of Variation Reduction on New Predictor

One would expect the decrease in B/W and Wc/Bc variability to enhance the ability to predict the %SV+SI in the samples. However, this did not prove to be true, with the predictive model actually decreasing in performance during the period of September through November. This loss of predictive performance indicates that there were sources of variation that were not being predicted by solely B/W and Wc/Bc. Since the new Hydrashear procedure looked very promising, little additional effort was made to find out the additional sources of variability.

In addition to finding a better predictor of project success, we tracked the effect of variation reduction program on output. Table 4 summarizes the output from library construction. The new predictor was placed in effect in September and the percentage of projects that passed Sequencing QC increased dramatically from historical (93% vs

75%). Unfortunately, many of the projects failed the "Overlap (O/L) Test "at that time, meaning that the project already overlapped an existing project, decreasing the number of projects actually delivered to Core Sequencing down to 27. In October, compounding the problem of overlap (only 42% of projects that made it through Sequencing QC passed the O/L test that month) was a marked decrease in percentage of projects that passed Plating and Sequencing QC. The Sequencing QC predictor was still working better than historical (June-August), but not as good as in September. The reason for the loss in predictive ability was not found, although there were a fair number of new vectors and reagents introduced that month. The library construction yields decreased further in November, with even less projects passing Sequencing QC. Library Construction had gone through a period of success followed by a period of failure. The exact reasons for these were not found during the internship, although the new hydrashear procedure promises to reduce variability significantly, which may help shed light on this subject in the future. As can be seen in the "November-H" period, which are the hydrashear projects done in November, the number of projects passing plating QC was 100% and only one out of 11 of these failed to make it through Sequencing QC.

Table 4 - Library Construction Statistics

Period	No. New Projects	No. Pass Plating QC	% Pass Plating QC	No. Pass Seq QC	% Pass Seq QC	No. Pass O/L Test	% Pass O/L Test	Overall Pass Rate	Seq QC Pass Rate
June	69	34	49%	26	76%	25	96%	36%	38%
July	55	37	67%	28	76%	23	82%	42%	51%
August	43	16	37%	6	38%	4	67%	9%	14%
Sept	73	43	59%	40	93%	27	68%	37%	55%
Oct	96	40	42%	29	78%	12	42%	16%	33%
Nov	61	26	43%	11	42%	11	100%	18%	18%
Nov -H	11	11	100%	10	91%	10	100%	91%	91%

C. Organizational Structure

1. The Challenge of Functional Organizations

According to Hayes, Wheelwright and Clark (HWC),[12] the modern business organization is based on two principles:

- Divide and Conquer – First enunciated by Julius Caeser two thousand years ago, where specialization is needed to be more efficient and line activities are separated from staff activities.
- Responsibility Equals Authority – A manager responsible for a certain scope should have the authority to commit the resources needed to accomplish the project.

The use of these principles creates a functional organization, where individual groups reporting to one manager are responsible for a particular function such as production, quality, maintenance, engineering, scheduling, materials management or purchasing. HWC argue that such an organization works well in a relatively stable business environment (5-10% productivity improvement per year), but when major improvements are needed simultaneously among several dimensions, the organization may not respond to well to such problems.

A general illustration of the functional problem occurs when the production group is made responsible by upper management to reduce costs. The production group may decide to lower costs by reducing the amount of preventive maintenance on their equipment. When the maintenance group is approached with such a proposal, they may resist because they are measured on cost of equipment failure, which they feel will eventually increase by suddenly downsizing PM program. In order for both sides to come to an agreement, they must both collaborate to share the risks and rewards, requiring an alignment that is difficult to negotiate if one party has asymmetric decision power.

[12] Dynamic Manufacturing, R. Hayes. S. C. Wheelwright and K. M. Clark, The Free Press, NY, 1988, pp. 96-129

2. Solving the Functional Organizational Problem

HWC proposes that a central staff can be used to get through the organizational barriers by auditing; evaluating performance; communicating goals and objectives; coordinating and prioritizing; consulting; management training; and advancing process development.

Another way of solving the problem is to adopt a process orientation to plant needs. For example, a plant may be organized such that a manager and the people reporting to her are responsible for a particular function (production, maintenance, etc.). A plant may also be organized such that a manager is in charge of a particular shift or group of people that does a particular process (it does not matter if they are all doing the same process or are on the same time shift). The advantages of process orientation is that it cuts across the organizational barriers, broadens manager's and worker's perspectives and allows for better learning between steps (similar to the advantages of Intel's "copy exactly" philosophy). However, the economies of scale must provide for the creation of distinct groups and the process must be relatively stable before doing this. The figure below illustrates the difference between process (horizontal shading) and functional organizations (vertical shading). Note that both types of organizations fulfill the two main principles of "Divide and Conquer" and "Responsibility Equals Authority".

Team/Function	Production	Maintenance	Eng/Dev	QA
Shift 1				
Shift 2				
Shift 3				
Shift 4				

3. Organization of the Center and Effects

At the beginning of the internship, the Sequencing Center was very functionally driven. As the internship progressed, the center moved to a more process-oriented management approach. The main places observed were the interactions between Library Construction,

Core Sequencing and Development. The Center had the following functional groups during the internship:

Mapping – conducting the main planning work of DNA sequencing including finding markers and preparing project-size DNA fragments that were roughly mapped.

Library Construction – taking the project DNA fragments, replicating, shearing, size selecting and packaging into recombinant form ready for core sequencing.

Core Sequencing – receiving the many small recombinant organisms that encompass a project processing and sequencing each one.

Finishing– obtaining the data from the Informatics group, identifying non-existent or questionable sequenced regions ("gaps" in the original DNA fragment) and doing laboratory work to prepare samples required to resolve problems.

Informatics – running both the IT infrastructure and software/hardware required to process and store the sequence data. Processing the data from the Core Sequencing and allowing access to the data by the different areas.

Materials Management and Infrastructure – staffing the stockroom, taking care of facilities, ordering supplies.

Development – executing new process or process improvement projects.

Each one of these groups had their own manager, usually a PhD or Masters level scientist specialized in a particular technology. There seemed to be alignment of objectives at the top and middle management levels, helped by the weekly staff meetings and daily discussions. As is observed in many manufacturing sites, there was less alignment of objectives and priorities at the factory floor level, due to the larger group size, more functional division and lack of forums at which to discuss and resolve problems.

An example of the central group working well together when the Center realized that a combination of M13's and pUC's gave the best sequencing results, with minimal number of gaps. The entire group was agile and flexible enough to convert to the new process in a matter of months.

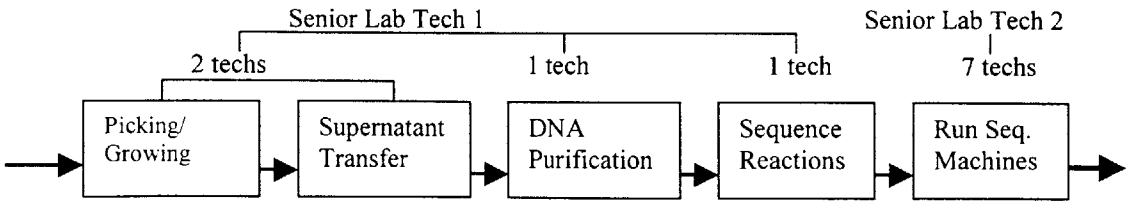
An example of organizational discord was the lack of balance in the types of projects that the development group was working on – they seemed to be top-driven, at first with little input from the core sequencing or library construction personnel, who were their main customers. The development group was in charge of specifying new equipment and building new automation and did so very efficiently. But many of their platforms had glitches which they did not have the time to fully address but caused a major amount of automation "babysitting" that core sequencing personnel had to do in order to quickly react to unexpected problems. As a result, the operations personnel could not just set up a batch and walk away from it, they had to stay pretty close to the machinery, limiting them from performing other tasks in parallel. Although this particular problem was a resource allocation issue, in that management deliberately wanted the group to focus on breakthrough projects, it nevertheless created some questions at the factory floor levels. The central question related to the balance between running fast and efficiently or versus working on the next-generation systems.

4. Assessment of Core Sequencing Organization

As a result of gathering data for the initial scale-up modeling for Core Sequencing, as well as form the results derived from the model, a number of potential areas of improvement were outlined.

In order to understand the context of the suggestions, the following summarizes the organization of Core Sequencing as it existed at the beginning of the internship:

The shotgun sequencing steps were organized hierarchically, with two groups of laboratory technicians, each headed by a more senior laboratory tech. The senior techs reported to a laboratory supervisor, who reported to the operations director. The responsibilities of each group are outlined below:



The main strategies for improving the factory floor personnel productivity included:

- Initiate cross-training. This came directly from suggestions outlined in the Manufacturing Scalability (meet gel machine cycle times), Inventory Management (minimize downtime by having more than one person able to run the automation) and Variation Reduction (standardize work and transfer knowledge) studies.
- Operate seven days a week, twenty four hours a day (7x24). In order to better utilize the existing infrastructure and to provide pressure to improve robustness of processes.
- Initiate team skills training. In order to go to a 7x24 operation, the teams would need to learn to rely on each other when management was not available to help them make decisions. In order to minimize problems due to poor team dynamics and additional "decisional" pressure imposed on the teams, the Center needs to invest in professional coaching and interpersonal skills training. There was also a technical component of skills training required to do continuous improvement work. The Center should invest in building the skills of their existing trainable personnel in order to improve their productivity.
- Change incentives and structure. The structure at the Center was modeled after the labs that Whitehead and MIT run. The hourly rate was relatively low, creating turnover of the most highly skilled personnel. The pay system was set up so people would get raises if they moved up the management ladder, creating horizontal layers and disenchantment due to the perceived limitation of upward mobility. Companies have found that paying for skills can become a powerful incentive for cross-training with even modest differences in pay. In organizations that require flexibility and technical skills, there are tradeoffs between having a few good people at a higher salary and having many people with low salaries.
- Training in process improvement work. The personnel at Core Sequencing had very little concept of learning curves, continuous improvement, and control charts. By actively involving them in making decisions about how to improve the process, they would feel more ownership and motivated to do their jobs better. This would also take some of the pressure off the development team to find improvements.

5. Outcomes of Core Sequencing Organizational Assessment

- The sequencing machine team started cross-training and evolved from having specific people wash and reload the gel plates to having everyone trained on the entire process. Further, they took on the tasks of gel-tracking, that is, taking the gel image output and checking/releasing the gel for further processing. As discussed in the previous sections, cross-training led to a reduction in variability and an increase in productivity, providing a good example of the benefits of moving a team from a functional to a process focus.
- The use of 7x24 shifts is still under consideration, but not implemented during the internship due to concerns about management supervision on weekends. However, some weekend work is nevertheless done unsupervised on an as-needed basis.
- Team skills training was not started because the Center decided to keep a hierarchical human resource policy.
- The Center changed the incentive structure for lab personnel by offering more competitive wages. Although the amount of pull from the industry at the time was not high, this could easily change if the local biotechnology companies continued to grow.
- Training in process improvement work was not done, since the development team was viewed as primarily responsible for improvement work. The perceived value of training lab personnel was low and any major suggestions would have to go to the development team to be implemented. However, "ad-hoc" improvement work was done by suggestions at the lab tech level that were brought to the attention of management. For example, in library construction, there was a suggestion to go to larger format plates, which was approved and implemented almost immediately.

D. Quality Assurance

Despite the desire to eliminate in-process testing by building the quality into the system, quality assurance (QA or QC) is a necessary evil in most processes. QA is used to avoid incurring additional processing costs on defective parts. Since it is usually not cost-effective to do 100% testing, a "batch" of the process can be sampled and if the quality is not deemed to be high enough, the entire batch is discarded. At the Whitehead Institute, there were many quality checkpoints that occurred at various parts of the process. We studied three of these in detail, to rationalize the threshold values and necessity.

1. Quality Assurance in the Core Sequencing Step

As discussed earlier, the bottleneck of the core sequencing step was the ABI (sequencing) machines. The Center's management desired to maintain a one-shift operation and to avoid adding personnel, even with a projected ramp-up to 40 machines. We also mentioned that quality control (QC) in the core sequencing step took up about 8% of the total labor costs (Exhibit 2-"QC Options"). These labor costs did not include quality control of the final product (gel tracking and approving). The first quality check step at the Shotgun Sequencing was after Supernatant Transfer and consisted of taking 16 samples out of each plate (96 clones) and running them through an agarose gel electrophoresis. The presence of DNA was detected by ethidium bromide, which in the presence of DNA, glows under UV light. The costs associated with doing the check included reagent costs, set-up cost for the gel, actually running the gel, taking a photograph, logging in results, discarding the gel and cleaning up the "gel box" for the next run. The mean failure rate at the supernatant transfer step was 1%, meaning that 1% of the samples would fail QC. An identical test was done after the DNA purification step, with a mean failure rate of 2%. An analysis using the Operating Curve concept [13] revealed that the Center could cut down their sampling to 8 per 96 wells with very little increase in Type 1 error (rejecting when the sample is acceptable), as shown in Exhibit 7. Exhibit 7 compares sampling 8 versus 16 wells. For example, given that the proportion

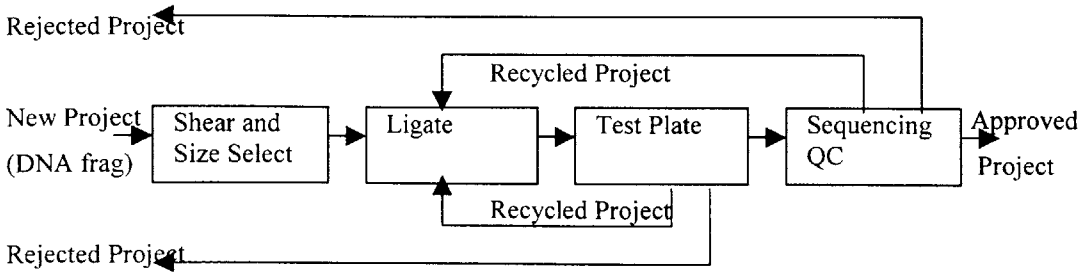
[13] "Production and Operation Analysis", Steven Nahmias, p.680, Irwin, 1997

of defects in the lot was 2%, the probability of accepting the lot is 99.97% (probability of rejecting is 0.03%) with 16 samples, whereas the probability is 98.86% (probability of rejecting is 1.14%) with 8 samples. The costs associated with the slight increase in Type 1 error was offset by the savings associated with less testing. Although net savings of $16,000 per year was projected, the Center could save as much as $24,000 per year by eliminating sampling altogether. However, since the sampling caught systematic errors that came up from time to time (for example a machine going out of calibration and this not being detected), it was deemed prudent to continue sampling at the lower level. Analysis revealed that the process would sometimes go out of statistical control and therefore this safety measure would catch the problems before they went on to the more expensive Sequencing reactions step. Since the automation platform was constantly being optimized for output coupled with a lack of a comprehensive preventative maintenance program (due to limitations in development resources), engineering the quality into the system was not considered at the time.

2. Quality Assurance in the Library Construction Step

As the Core Sequencing step started ramping up in scale, the Library Construction step was quickly becoming the new bottleneck of the Center. Further, due to the challenge to the project by private investors (Perkin-Elmer's Celera), the need for finished projects became less pressing, while the need for sequence data to get a "rough" version of the Human Genome increased. The following flowchart shows the main steps in Library Construction:

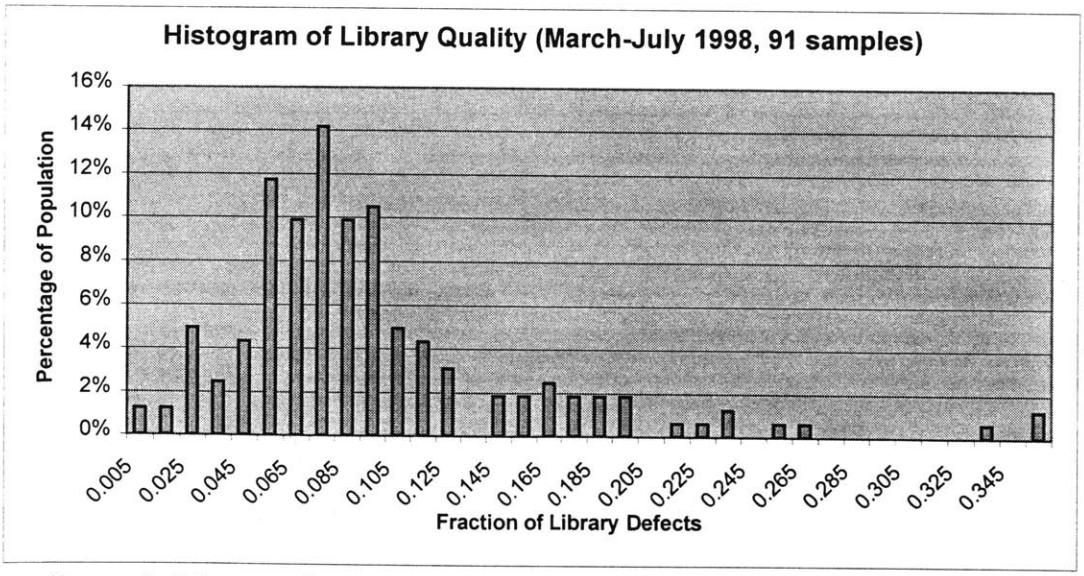
Library Construction QC Process:



As discussed in the Variation Reduction chapter, once a project passes the Plating QC step, the colonies (plaques) go on to a more elaborate test, where it is actually run through the Core Sequencing step as a trial. The Sequencing QC step tests for three things: Empty Vector (meaning no DNA inserted into the vector), Small Insert (very small pieces of DNA inserted into the vector) and *E. coli* (meaning DNA from *E.coli* rather than human was sequenced). If the aggregate amount of defects for the test plates exceeds a certain threshold (8% at the beginning of this process), the project is rejected and the library construction process must be started over again. We found that the rate of project rejection at the Sequencing QC step would be about 45% if the criteria of 8% were seriously followed. However, in an effort to keep the Core Sequencing running, exceptions were made and projects were passed that were above this threshold, resulting in about 75% of the projects passed. In order to find the true most cost-effective threshold value, we decided to build a cost-based model (shown in Exhibit 8).

Cost-based model of the Sequencing QC step

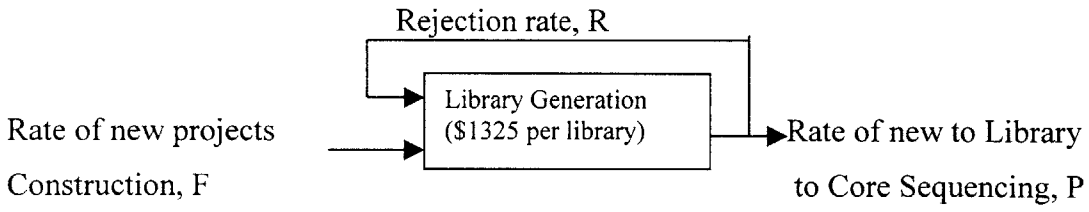
The premise of the model (Exhibit 8) was to determine the cost of library quality. Library



quality varied from project-to-project, as shown by the histogram above. The cost of a library could be broken down into the cost of generating a library and the cost of lost data due to low library quality.

Ideally, a project was generated at an incremental cost of $1325 (labor, reagents). Even if this project passed Sequencing QC, as outlined above, a certain fraction of its data, D, would be unusable due to defects associated with library construction. The cost of sequencing a plate is $225 each and each project has about 30 plates. Therefore, the cost of library generation and lost sequencing data for D=0.10 is:
$1325 +$225*30*0.10= $2000, where D is the fraction of the project

However, one can reduce the cost of lost data by screening out libraries above a given threshold, which has the effect of producing libraries of higher quality. An example of this is shown above from a sample of libraries taken from a given period in time. The mean of the distribution shown above is 0.10 (10% defects). If all of the libraries with defects rates above 11% (0.11 fraction) are screened out, the mean of the truncated distribution drops from 0.10 to 0.08.
However, screening will cause more libraries to be rejected and it is shown below that the cost of library generation will increase due to "recycling" of projects within Library Construction:



At steady-state, F = P, the rate of new libraries introduced into the system equals the rate that leave the system.
F+R = L, library generation rate, the rate that libraries must be generated to account for both the new ones and reprocessing of the existing ones.
Let r = R/(R+F), the fraction of libraries that are rejected.
Therefore R= r (R+F) or
R(1-r) = r F or
R = F(r/(1-r))
Therefore, L = F(1+r/(1-r))

Therefore, for a rejection rate of 5% that gives a mean project defect rate of 8% (from the sample given above),

L= 1(1+0.05/(1-0.05))=1.053

The costs to generate one library becomes:

$1325*(1.053)+$225*30*0.08=$1395+540=$1935.

One of the key functions of the equation is the rejection rate due to the threshold used. If the same screening threshold gives a 10% rejection rate, then the costs become

$1325*(1.111)+$225*30*0.08=$1472+$540=$2012.

The last part of the cost model accounts for the fact that if the data from sampling was completely reusable, a project could completely recover the costs of the extra sequencing required by rejected samples by simply running a lower "coverage" rate. Often, a library that is rejected has such poor data that it is not used in the final assembly process. To account for this loss of data, the model assumes that once a library is rejected, only about 50% of its data are recoverable, which means that the project has incurred additional costs due to lost data. The lost data costs can be estimated by R*$225*50%. Since R =F(r/(1-r)), on a unit cost basis (F=1), the lost data cost is:

$225*0.053*0.5=$6 for the second example and $225*0.11*0.5=$12 for the third example.

The three examples give values of $2000 (taking the distribution as given), $1941 (11% cutoff and 5% rejection) and $2024 (11% cutoff and 10% rejection). Clearly, there is some value in going through the exercise of finding the optimum threshold value of rejection.

The threshold used for rejection gives different rejection rates depending on the sample size. Using binomial theory, one can estimate the rejection rate at a given mean value of the sample. All of the factors discussed above are shown in Exhibit 8 for three different sample sizes. We used Crystal Ball to set up a Monte-Carlo simulation of these events assuming a fitted log-normal distribution of library failures.

Outcome

The results from the simulation appear in Exhibit 8, showing that costs increase significantly as the threshold value of is decreased. However, there are diminishing

returns to increasing the maximum acceptable levels. Having no limit actually is problematic, in that libraries with high percentages of defects create assembly and finishing problems not represented in the model. In particular, data assembly problems, shortage of coverage for projects and added logistical complexity start becoming a large factor at library qualities with higher than 15% defects. The model also validate thoughts that there should be three test plates per QC, showing a sharper drop to the minimum cost level, although higher costs if the threshold value is low. The Center decided to use 12% as the new, official threshold, theoretically dropping their Library Construction costs from $3686 to $2137 per library. At the high levels of throughput (50 libraries per month), this could result in savings of up to $929,400 per year. However, most of the time the 8% rule was already being by-passed.

E. Inventory Management

Motivations Behind Holding Inventory

The classical reasons for holding inventories are[14]:

1. Uncertainty of supply and demand
2. Setup costs
3. Speculation
4. Transportation
5. Smoothing
6. Logistics
7. Control Costs

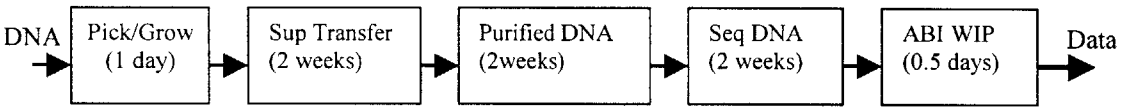
Although these motivations are relevant to different extents, there is an additional motivation not mentioned in Nahmias, which is to balance throughput with development speed. There is a need to maintain a minimum amount of inventory in order to avoid shortages and subsequent downtime. However, if the amount of inventory is too high, it increases the ramp-up time for development efforts. This concept is one of the reasons the JIT system is successful, it allows quick identification of problems (cause and effect) due to less lag time in identifying and minimizes the amount of defective inventory when a process problem occurs. However, implementing JIT requires a large amount of management and supplier commitment. At this stage of manufacturing capacity development, it was clear that JIT would be difficult to implement correctly. Therefore, we decided to look at inventory levels at the Core Sequencing step in order to find an optimum.

Core Sequencing Inventory Model

The Core Sequencing step had various places where work-in-process inventory (WIP) was accumulated. Further, each processing step had different mean daily capacities and variances. The daily capacities were influenced by raw material availability, staffing and

[14] "Production and Operation Analysis", Steven Nahmias, p.213, Irwin, 1997

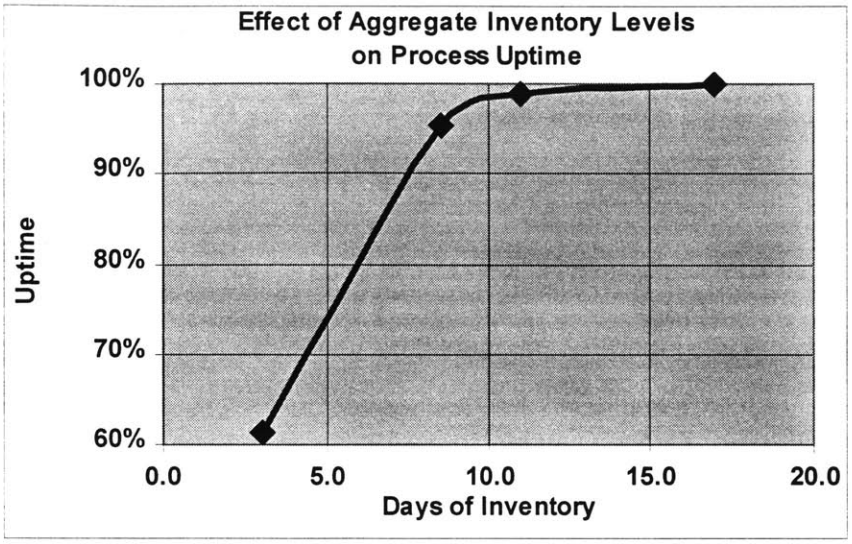
machine uptime. The first four steps were particularly vulnerable to absenteeism, due to the specialization of the labor functions. As a result, the lead technician tended to keep an inventory of two weeks of production between steps, adding up to six weeks of inventory in the system. Correspondingly, any new process improvements would take at least that amount of time to make its way through the system. In order to minimize the time to test new processes, the inventory (FIFO system) was bypassed on an as-needed basis. However, full implementation still required eventual replacement of the entire inventory.



The model used data obtained from the current operations to determine the throughput pattern of each step (Exhibits 9-12). Each of the Exhibits 9-12 has a 110-day sample of daily throughput data for the particular step. The data were then sorted as a histogram (shown on each exhibit) in order to simulate the variability on output for each step. Based on the histogram, assuming that the distribution will not change, we built a stochastic model and ran simulations with varying days of inventory at each step. The model (Exhibit 13) assumes that each particular step has a max-min inventory level. The step would either run or not run depending on the inventory level. Each step was assumed to be capable of its design throughput. The actual throughput of the step would depend on a "multiplier", which had the pattern from the original histograms imbedded. Therefore, even if a step "ran" on a particular day (due to low inventory levels), its throughput would be random and bounded by the histogram pattern. The output for the entire process was linked to the variability of throughput of each individual step. At any given time, if its raw material inventory levels dropped below zero, the final step's output could be zero. The monthly output could be characterized as a percentage of theoretical (uptime).

The results of the simulation are shown below, where as expected, the percent uptime increases with inventory, to a point of diminishing returns. For this particular process,

the 99% uptime was set at 11 days of inventory, compared to the current practice of 30 days of inventory.



Outcome

The inventory reduction model was used to point out that the optimum inventory was far below that used in practice. Although the levels of inventory were not formally controlled, there was greater awareness of the effect of having large inventories on the speed of development, which was an important part of the ramp-up process. After this model was developed, inventory levels dropped due to low input rates from Library Construction, so inventory levels remained low. Future plans called for the next-generation automation system to replace many of the labor-dependent operations that create the variation in output for each process. However, it was also noted that by cross-training personnel on Core Sequencing, the group would have more flexibility to shift resources to the critical step as required.

IV. Process Development Decision Support Hypothesis

A. Hypothesis

The hypothesis is that high variation processes are more conducive to improvement through radical improvement efforts rather than incremental. The variability of processes can be described by their outputs using the sample coefficient of variation (Sx/X) to classify them. If a process is classified as high variance, the resources required to do incremental improvement are better utilized on radical improvement efforts, assuming that basic scale-up data are available. If a process is classified as low variance, the resources may be better spent on incremental improvement to take advantage of learning curve effects, assuming that scale-up data are available. These decisions are particularly important in biological processes of today, where a high variance process may be feeding a low-variance process and decisions must be made on what projects to work on with limited development resources.

B. Theory

The sample coefficient of variation, defined as Cv is simply the ratio of the standard deviation to the mean:

Cv =Sx / X, where C is the coefficient of variation, Sx is the sample standard deviation and X is the sample mean.

The higher the coefficient of variation, the more difficult it is to discern incremental improvements of the sample mean. The difficulty may be overcome by taking many measurements of the effects of small changes on the process over time, also called evolutionary operation (EVOP). Depending on the rate of generation of new measurements, a modest process improvement may take a long time before it can be confirmed as having had an effect. As discussed in the Variation Reduction chapter, the Core Sequencing process had a relatively low coefficient of variation, less than 0.1 (Table 2). In contrast, Library Construction had a relatively high coefficient of variation, 0.5-1 (Table 3). Clearly, it is easier to discern small differences in performance in Core Sequencing than in Library Construction.

C. Results

The Center applied many incremental improvement methods to Core Sequencing, resulting in high improvements in performance. When similar incremental improvement methods were applied to Library Construction, there was a temporary increase in performance initially, followed by a return to previous levels. Although there may have been extenuating circumstances, the application of incremental tools did not correlate well with increased performance. The reason for this is because it is very difficult to do incremental improvements in the face of such high variability. Although some of the variability was decreased, it was still not low enough to quickly find process improvement paths. Additionally, there were special circumstances in Library Construction, such as all tasks being performed manually with some degree of lab skill required and high personnel turnover. In contrast, Core Sequencing was more fully automated for the process critical tasks and had the full attention of the development group.

During the internship, operations director decided to undergo a parallel improvement path, where the development group was asked to work on an independent, radical improvement effort designed to replace the existing Library Construction technology. There were many changes in Library Construction; one of the large ones was in the shearing operation, which was switched from an ultrasonic shearing to a "point sink" or "hydroshearing" technique. According to authors of this new technique[15], the coefficient of variation is less than 0.1. By the end of the internship, the technique was successfully scaled up and implemented in Library Construction, resulting in a radical improvement in their ability to supply the Core Sequencing step.

D. Conclusions

The ability to quickly ascertain the viability of incremental improvement efforts is an utmost priority in the manufacturing sector and a simple statistic such as the coefficient

[15] "An Automated Hydrodynamic Process for Controlled, Unbiased DNA Shearing", Genome Research, Y. Thorstenson et al, Vol. 8, Issue 8, pp. 848-855, August 1998.

of variation can be used to help in decision-making. Although we have not addressed the "middle ranges" of the coefficient of variation, it is clear that if a process is at either end of the spectrum, the decision should be relatively straightforward.

The next question addresses the situation when there is no scale-up data available for radical improvement. At this point, technical judgement should be used to decide whether to try the incremental process using more appropriate techniques such as EVOP or design of experiments (DOE) and/or generate scale-up data for radical improvement

V. Conclusions

As the process for DNA sequencing matures, the people who run these processes must decide whether to improve their processes by incremental improvement (staying on the S-curve) or by radical improvement (moving to the next generation S-curve). The tradeoff is the balance between taking advantage of the learning curve associated with continuous improvement versus the disruptive effects of a large step change improvement. The decision process is made even more difficult when the challenge is to increase output by factors of three or more per year. To add another layer of complication, biological processes are notoriously high in variation, making it even harder to decide which way to go. The Center for Genome Research has done a remarkable job of balancing the tradeoffs associated with these decisions, as witnessed by their recent increase in output and subsequent extension of funding by the NIH.

A. The Use of Manufacturing Management Tools

Although the use of manufacturing management tools provide insights and can be very valuable in the decision making process, they must be directly applicable to the problem and be part of an overall operational strategy. In order to make the use of manufacturing tools more relevant, they must be customized and evaluated for usefulness, requiring a combination of technical expertise, experience and judgement. A high efficiency organization must build their own tools to order to best evaluate and recommend alternatives.

Manufacturing tools by themselves yield some marginal results, but work best when they are part of the overall operational strategy and are applied where synergies can occur. An example of this is cross-training: By cross training, the degree of variability eventually decreases due to knowledge transfer, the ability to more easily respond to down-time lowers the need for inventory. These effects lead to increased productivity and employee engagement, eventually increasing the demand for cross-training. Good operational

strategy requires a self-reinforcing, self-perpetuating system (virtuous cycle as defined by Peter M. Senge[16]) where the needs of the organization and the people running the organization are met.

Focusing a research organization on process improvements means changing a fundamental part of their culture. As discussed by Schein,[17] cultures are very stable and changing culture in any organizations is very difficult, requiring strong sustained leadership and commitment at all levels. During the internship, the Center showed leadership at the management level, but not necessarily at the worker level. It seems like there is some resistance at the lower levels, mostly due to their perceived lack of participation in improving the process.

B. Final Recommendations

Some additional virtuous cycles that the Center may be able to take advantage of include:

- Monitor variation and reduce it – although the Center prides itself in having much sequencing data available on-line, it would be to their advantage to find some variation statistics they can monitor to help them in finding ways of reducing variability in the process.

- Improve the information exchange. At the factory-floor level, there is a lack of metrics associated with targets. Without the right metrics and association with the targets, it is difficult for employees to help improve the process. Although the overall Center goals are clear, there is too much of a lag between what is done at the factory floor and its effect on the process. The high variability of the processes emphasizes the need for tools such as control charts to help differentiate a change in the process from its normal variability.

[16] "The Fifth Discipline: The Art and Practice of the Learning Organization", Peter M. Senge, Doubleday/Currency, 1990
[17] "Organization Culture and Leadership", Edgar H. Schein, p.298, Jossey-Bass, 1992

- Carefully consider use of QA – continuously monitor or eliminate the need for QA, invest in making the process more robust rather than relying on testing.

- Manage inventory more carefully – as mentioned throughout the thesis, use low inventory levels to help uncover erratic processes but mostly to reduce new development ramp-up time.

- Increase employee involvement – enable the factory floor people to participate in continuous improvement, assign a small workforce to do troubleshooting exclusively for them or train some of them to do the routine maintenance replacement.

- Much of the incremental improvement shown in this process can be attributed to following the principles of cGMP such as change control, detailed records, validation of effects of changes on process and clear procedures. Although it may seem onerous and costly to have a research organization follow cGMP/cGLP, the discipline imposed may help the organization take better advantage of the learning curve.

VI. Appendix

List of Exhibits:

- Exhibit 1 (4 pages):
  Whitehead Institute Operational Summary – Core Sequencing Operations
- Exhibit 2 (4 pages – B size)
  Whitehead Institute Process Flow Diagram- Core Sequencing Operations
- Exhibit 3 (4 pages)
  Whitehead Institute Scheduling Summary - Core Sequencing Operations
- Exhibit 4 (8 pages)
  Whitehead Institute Statistical Measurements - Core Sequencing Operations
- Exhibit 5 (1 page)
  Regression of B/W(Wc/Bc) Vs. % SV
- Exhibit 6 (1 page)
  Regression of B/W(Wc/Bc) Vs. % SV+SI
- Exhibit 7 (1 page)
  Operating Characteristic Curves
- Exhibit 8 (2 pages)
  Whitehead Institute Library Construction Sequencing QC
- Exhibit 9 (1 page)
  Whitehead Institute Histogram of Picked Plates Per Day
- Exhibit 10 (1 page)
  Whitehead Institute Histogram of Purified Plates Per Day
- Exhibit 11 (1 page)
  Whitehead Institute Histogram of Sequenced Plates Per Day
- Exhibit 12 (1 page)
  Whitehead Institute Histogram of Plates Run on ABI Per Day
- Exhibit 13 (1 page)
  Whitehead Institute Stochastic Simulation of Inventories for Core Sequencing

Exhibit 1: Whitehead Institute Operational Summary - Core Sequencing Operations (June, 1998)

Scenario:

76.0 plates/day
40 ABI machines
29 plates/project

Throughput:

Average Project Size:	29	plates	Project Plate Breakdown:		Proj Alloc
Plates/day:	76.0		SS's=	12 plates	41.4%
Days/week:	**5**		DS's=	12 plates	41.4%
Weeks/year:	**52**		QC/Lib=	**2** plates	6.9%
Projects/year:	681		Finishing=	**2** plates	6.9%
Average Project Size:	**100,000**	base pairs	End,mkrs=	**1** plates	3.4%
On-line Factor:	90%		Total=	29 plates/project	
Bases/year:	61.3	MB			100.0%

Shotgun coverage pattern:

	SS	DS	Totals	SG plates
FP	**8**	**6**	14	58.3%
RP	**0**	**6**	6	25.0%
FT	**4**	**0**	4	16.7%
RT	**0**	**0**	0	0.0%
Totals:	12	12	24	

This part of the spreadsheet is used to establish the throughput assuming the ABI Sequencing machines are the bottleneck. The shaded areas are variables that can be changed to study effects.

ABI Machines:

	No. ABI's	lanes/gel	plates/gel	ABI cycle (hrs)	gels/day	Project plates/day	Proj. Alloc
Shotgun	**31.4**	**96.0**	1.00	**12**	62.9	62.9	82.8%
QC/Library	**2.6**	**96.0**	1.00	**12**	5.2	5.2	6.9%
Finishing	**2.6**	**96.0**	1.00	**12**	5.2	5.2	6.9%
End,mkrs	**1.3**	**96.0**	1.00	**12**	2.6	2.6	3.4%
Dev/Down	**2.0**	**96.0**	1.00	**12**	4.0		
Total ABI's	**40.0**		1.00		80	76.0	100.0%

Exhibit 1 (continued): Whitehead Institute Operational Summary - Core Sequencing Operations (June, 1998)

Automation Factors:

Purpose:	SS Picking	S Sup Xfer	S Sup Xfer	SS Purif	DS Purif	SS/DS FP	SS/DS RP	SS/DS FT	SS/DS RT
Machine:	Flexys	Packard	Packard	Tecans	Hydras	Sequatron	Sequatron	Cascade	Cascade
No. machines	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
No. of steps/machine	**4**	**1**	**1**	**10**	**15**	**5**	**5**	**2**	**2**
Machine cycle time	**10.0**	**0.6**	**0.6**	**6.7**	**6.0**	**2.5**	**2.5**	**2**	**2**
Max plates/machine	**8**	**8**	**8**	**15**	**40**	**20**	**20**	**8**	**8**
Batch size (plates)	8.0	8.0	8.0	15.0	16.0	19.0	16.0	8.0	0.0
Set-up time/batch	**10**	**2**	**2**	**10**	**10**	**5**	**5**	**5**	**5**
Process time/batch	110.0	4.67	4.67	160.0	180.0	57.50	50.00	18.00	4.00
Cycle time/batch	120.0	6.67	6.67	170.0	190.0	62.50	55.00	23.00	9.00
Labor time/plate	1.3	0.4	0.4	6.0	6.3	1.0	1.1	1.1	#DIV/0!
Residence time/plate	120.0	6.67	6.67	170.0	190.0	62.50	55.00	23.00	9.00
"Baby sit" factor	0%	25%	25%	50%	50%	25%	25%	20%	20%

Purpose:	FP TC	RP TC	FT TC	RT TC	Pool P	Cascade T
Machine:	Tetrads	Tetrads	Tetrads	Tetrads	P Tecan	Cascade
No. machines	**5**	**5**	**5**	**5**	**1**	**1**
No. of steps/machine	**1**	**1**	**1**	**1**	**1**	**1**
Machine cycle time	**60**	**60**	**150**	**150**	**4.5**	**2**
Max plates/machine	**4**	**4**	**4**	**4**	**1**	**8**
Batch size (plates)	19.0	16.0	2.0	0.0	1.0	8.0
Set-up time/batch	**5**	**5**	**5**	**5**	**0.5**	**5**
Process time/batch	60.00	60.00	150.00	150.00	4.50	16.00
Cycle time/batch	65.00	65.00	155.00	155.00	5.00	21.00
Labor time/plate	0.3	0.3	2.5	#DIV/0!	0.5	1.0
Residence time/plate	65.00	65.00	155.00	155.00	5.00	21.00
"Baby sit" factor	0%	0%	0%	0%	0%	20%

▲ **This part of the spreadsheet is used to enter data regarding automation and labor associated with it. The shaded areas are variables that can be changed to study effects.**

Exhibit 1 (continued): Whitehead Institute Operational Summary - Core Sequencing Operations (June, 1998)

Labor Factors:

Purpose:	# Shifts	Req'd labor per day (hrs)	% Total	Assigned labor per day (hrs)	% Usage	People assigned	Plates per day	Batches per shift	Batch size (plates)	Max Batch size (plates)
SS Picking	**1**	10.02	8%	11	91%	1.38	31.4	3.93	8.0	**8**
DS Picking	**1**	5.01	4%	6	84%	0.75	15.7	1.97	8.0	**8**
SS Sup Xfer	**1**	3.30	3%	4	83%	0.50	31.4	3.93	8.0	**8**
DS Sup Xfer	**1**	1.68	1%	2	84%	0.25	15.7	1.97	8.0	**8**
SS Sup QC	**1**	2.78	2%	3	93%	0.38	31.4	1.05	30.0	**30**
SS Purif	**1**	6.36	5%	7	91%	0.88	31.4	2.10	15.0	**15**
SS Purific QC	**1**	3.98	3%	4	100%	0.50	31.4	2.10	15.0	**15**
DS Purific	**1**	2.54	2%	3	85%	0.38	15.7	0.98	16.0	**16**
DS Purific QC	**1**	2.74	2%	3	91%	0.38	15.7	0.98	16.0	**16**
SS/DS FP	**1**	8.69	7%	9	97%	1.13	36.7	1.93	19.0	**19**
SS/DS RP	**1**	5.23	4%	6	87%	0.75	15.7	0.98	16.0	**16**
SS/DS FT	**1**	5.28	4%	6	88%	0.75	10.5	1.31	8.0	**8**
SS/DS RT	**1**	0.00	0%	0	n/a	0.00	0.0	0.00	0.0	**8**
ABI load/prep	**2**	27.12	22%	28	97%	3.50	76.0	**4.00**	9.5	**10**
ABI plate prep	**2**	25.38	20%	26	98%	3.25	76.0	**4.00**	9.5	**10**
Coordinators	**1**	16.00	13%	16	100%	2.00				
Total		126.12	100%	134	94%	16.75				
Min # people req:		15.76								

▲ **This part of the spreadsheet is used to enter data regarding shifts and calculates number of batches per shift needed. The shaded areas are variables that can be changed to study effects.**

Labor Break-Out

Task	Labor	% Labor
QC (all)	9.50	8%
Picking (all)	15.04	13%
Sup Xfer (all)	4.98	4%
SS Pur	6.36	5%
DS Pur	2.54	2%
Primer Seq	13.92	12%
Term Seq	5.28	4%
Coordination	8.00	7%
ABI load/prep	27.12	23%
ABI gel	25.38	21%
Totals	118.12	100%

Processing Factors:

Step	Residence Time (hrs)	Inventory Time (hrs)	% time in process (hrs)
Pick/Xfer	19.3	72	21.2%
Purification	4.1	72	5.4%
Sequencing	2.8	72	3.7%
Gel running	13.1	0	100.0%
Total	39.30	216.00	18.2%

Exhibit 1 (continued): Whitehead Institute Operational Summary - Core Sequencing Operations (June, 1998)

QC Options (sup, purif):

Scenario	Sup QC (wells/plate)	M13 QC (wells/plate)	PUC QC (wells/plate)	Capital Cost/year	Reagent Cost/year	QC labor Cost/year	Prod. Loss Cost/year	Net Cost/year
Current	16	16	10	$ -	$ -	$ -	$ -	$ -
Cytofluor	16	16	10	$ 8,000	$ 47,697	(55,556)	$ -	$ 141
0/16/10	0	16	10	$ -	$ -	(18,057)	$ 7,849	$(10,207)
0/8/10	0	8	10	$ -	$ -	(31,003)	$ 15,699	$(15,304)
0/0/10	0	0	10	$ -	$ -	(43,949)	$ 23,548	$(20,400)
16/8/10	16	8	10	$ -	$ -	(12,946)	$ 7,849	$ (5,097)

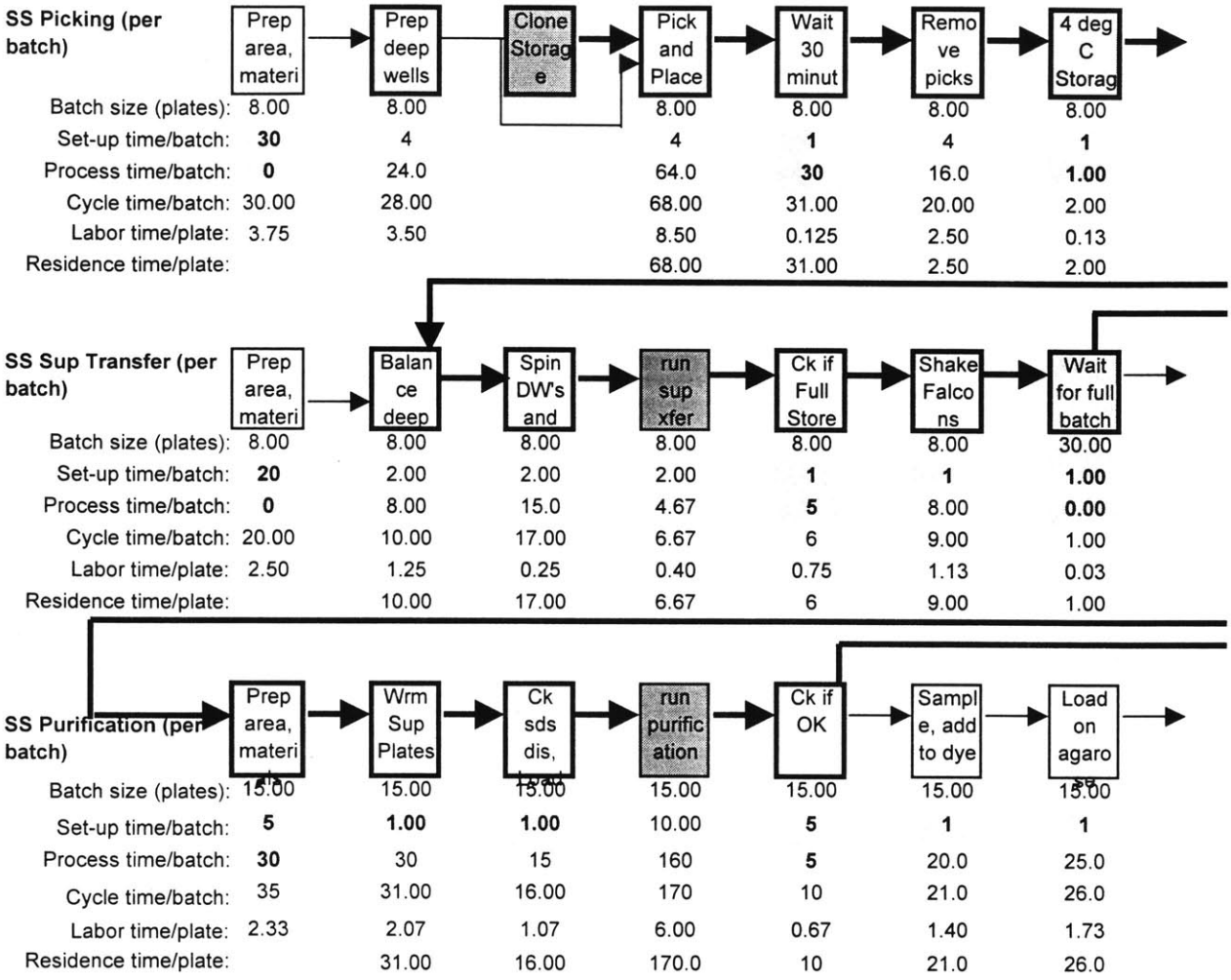
Capital cost:	$ 16,000		
Depreciation period:	2	years	
Overall plate "fail rate"=	3.0%		
Labor cost:	$ 25	per hour- based on base 11, fully loaded	
$/plate=	$ 48	assumes no fixed head count	

	Plates/day	Fail rate
Sup Xfer	40.0	1%
SS purifica	31.7	2%
SS seq/AB	31.0	0%
DS purifica	7.9	2%
DS seq/AB	7.8	0%
ABI gel	38.8	3.0%

Other issues: Cytoflour reagent costs may be lower through dilution, may not give "extra" data that gel gives.

▲ **This part of the spreadsheet is used to compare various QC options for the supernatant transfer and purification steps. The shaded areas are variables that can be changed to study effects.**

Exhibit 2: Whitehead Institute Process Flow Diagram - Core Sequencing Operations (June, 1998)



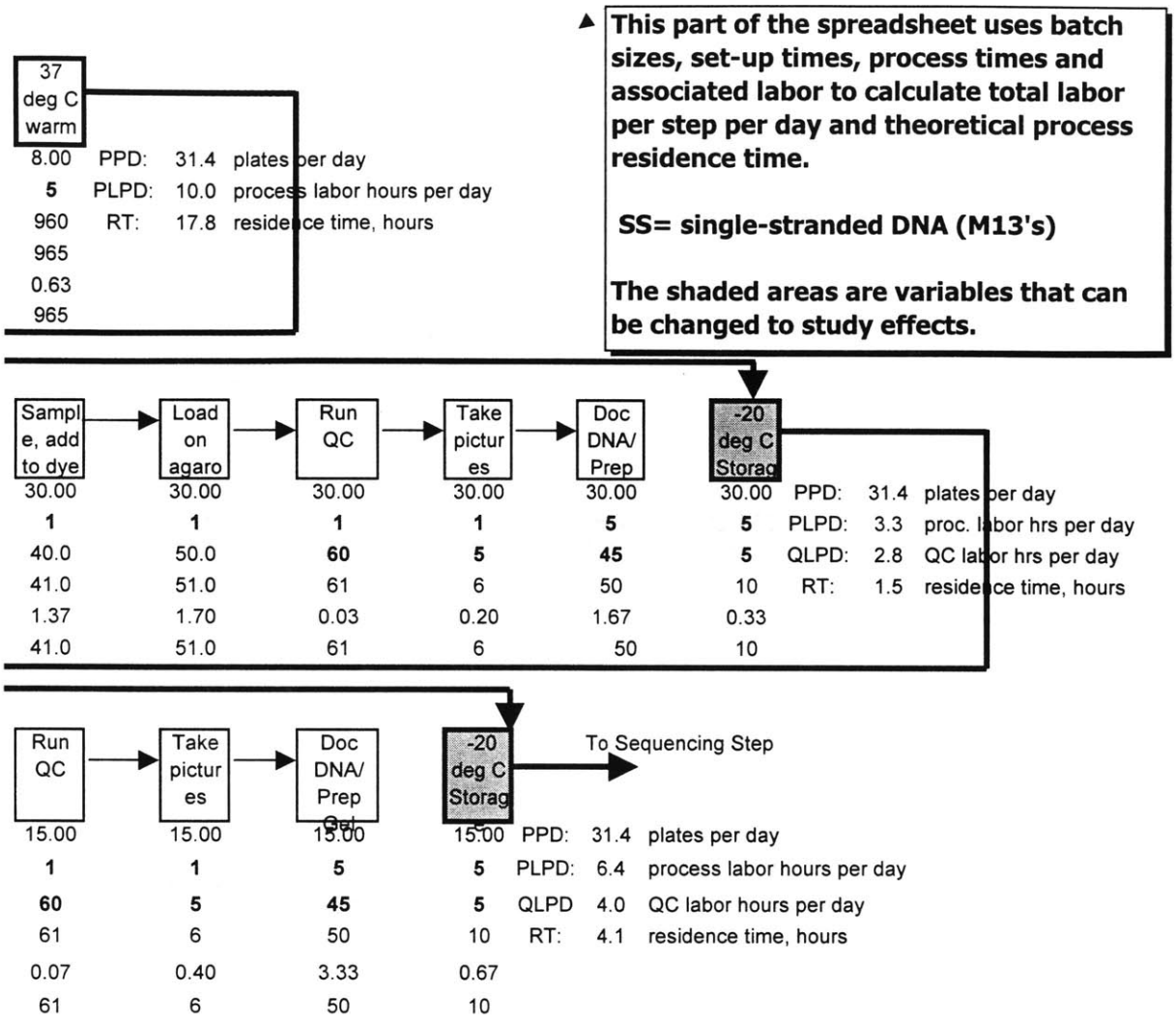
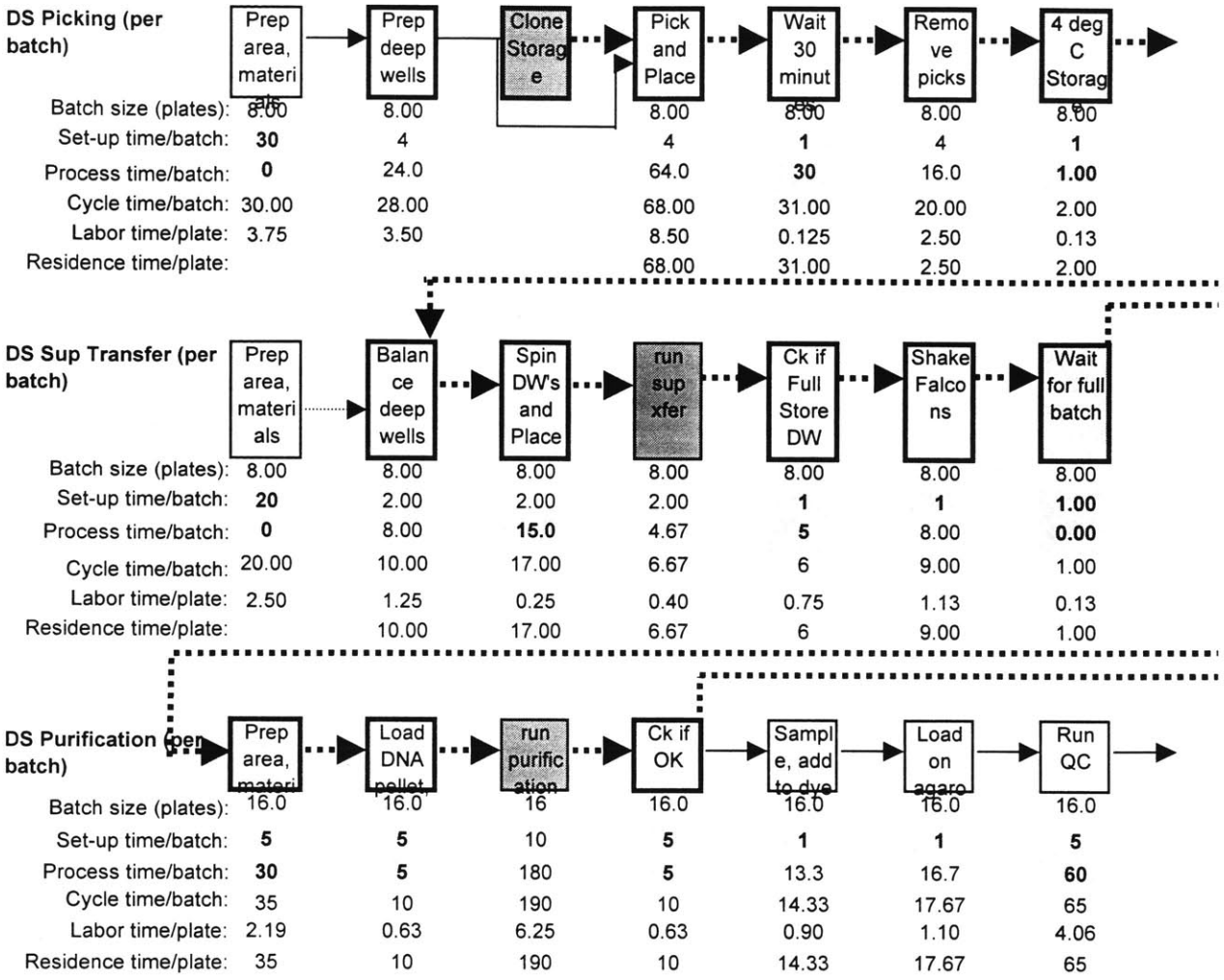

This part of the spreadsheet uses batch sizes, set-up times, process times and associated labor to calculate total labor per step per day and theoretical process residence time.
SS= single-stranded DNA (M13's)
The shaded areas are variables that can be changed to study effects.
37 deg C warm
8.00 PPD: 31.4 plates per day
5 PLPD: 10.0 process labor hours per day
960 RT: 17.8 residence time, hours
965
0.63
965
Sampl e, add to dye
Load on agaro
Run QC
Take pictur es
Doc DNA/ Prep
-20 deg C Storag
30.00 30.00 30.00 30.00 30.00 30.00 PPD: 31.4 plates per day
1 1 1 1 5 5 PLPD: 3.3 proc. labor hrs per day
40.0 50.0 60 5 45 5 QLPD: 2.8 QC labor hrs per day
41.0 51.0 61 6 50 10 RT: 1.5 residence time, hours
1.37 1.70 0.03 0.20 1.67 0.33
41.0 51.0 61 6 50 10
Run QC
Take pictur es
Doc DNA/ Prep Gel
-20 deg C Storag
To Sequencing Step
15.00 15.00 15.00 15.00 PPD: 31.4 plates per day
1 1 5 5 PLPD: 6.4 process labor hours per day
60 5 45 5 QLPD 4.0 QC labor hours per day
61 6 50 10 RT: 4.1 residence time, hours
0.07 0.40 3.33 0.67
61 6 50 10

Exhibit 2 (continued): Whitehead Institute Process Flow Diagram - Core Sequencing Operations



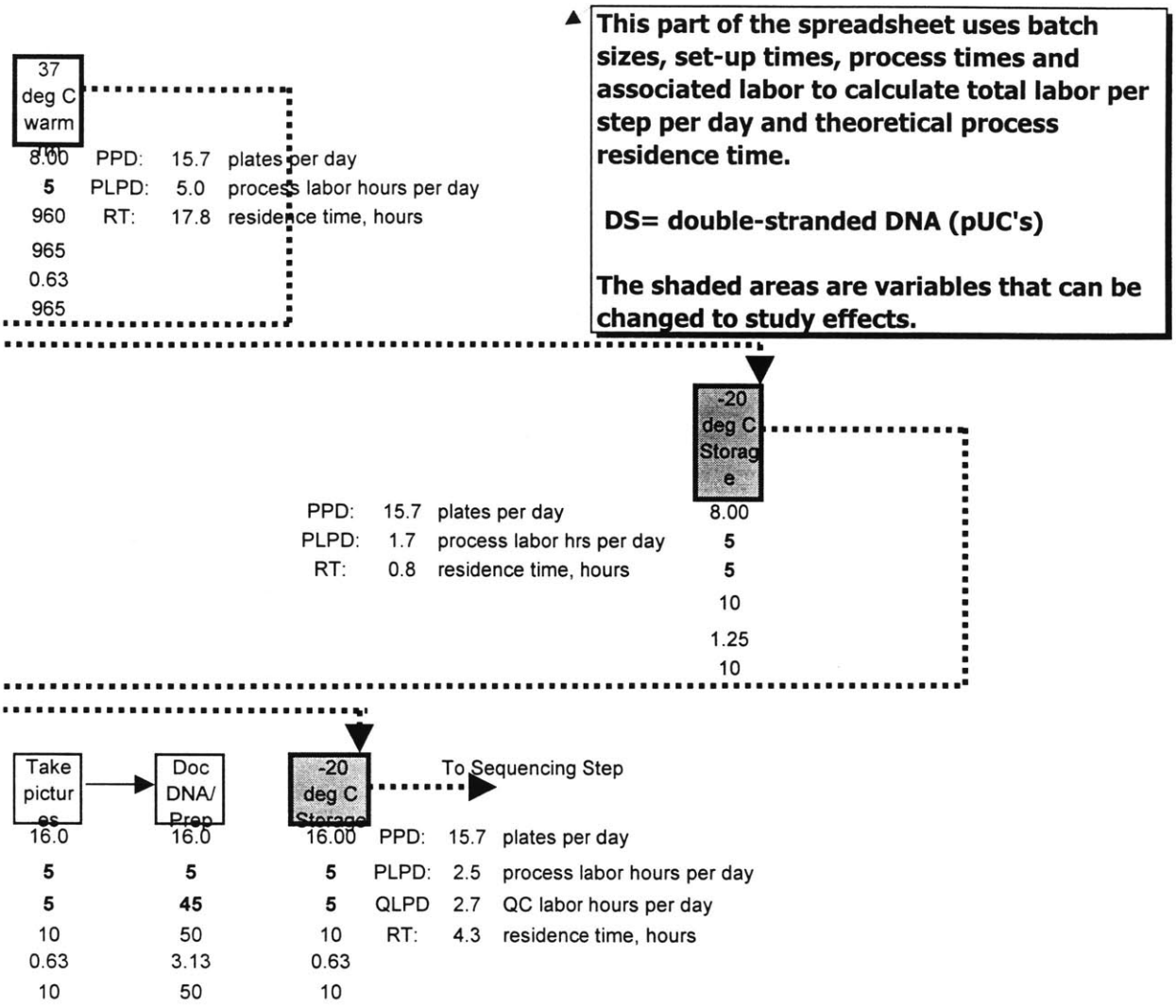
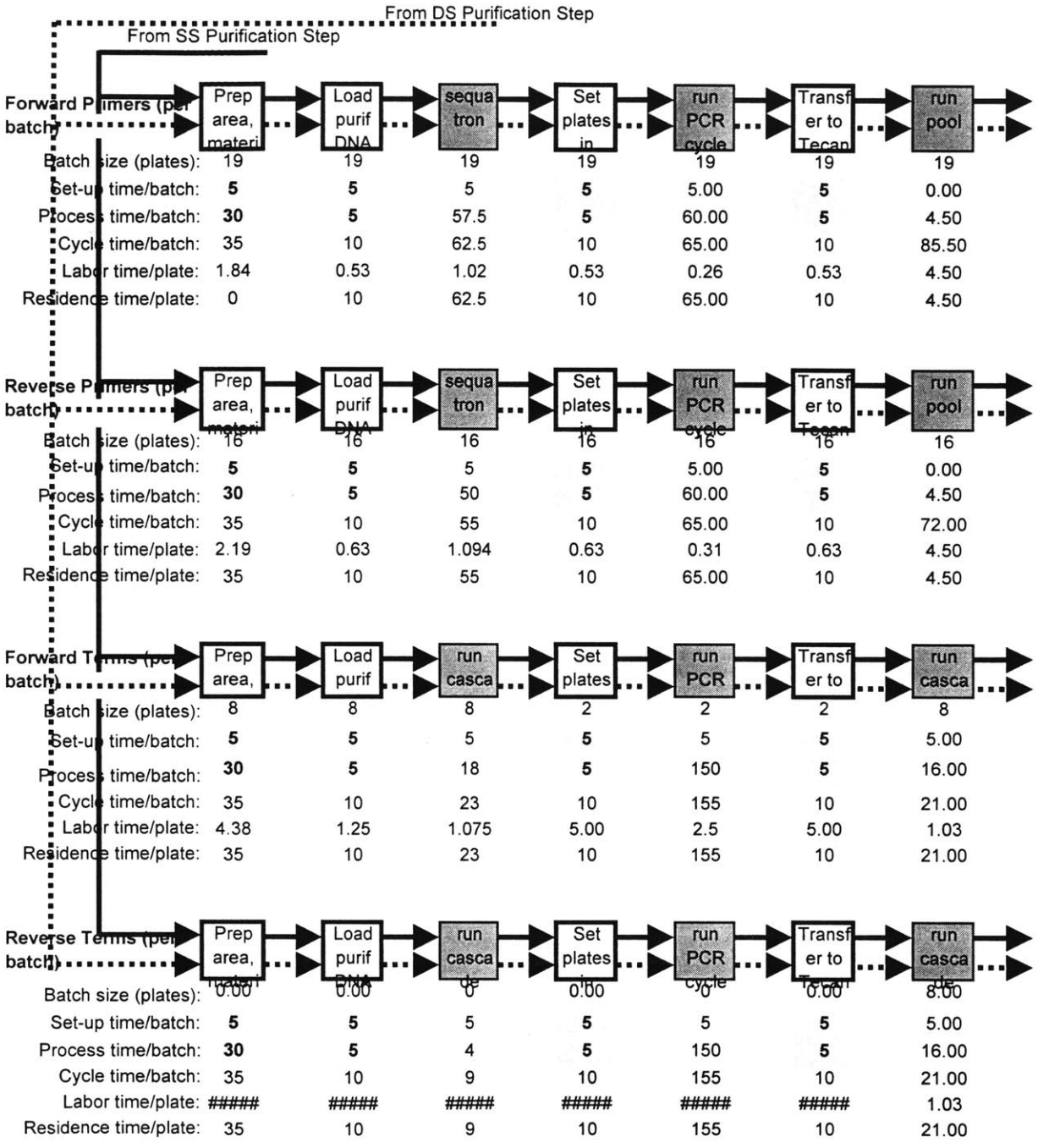
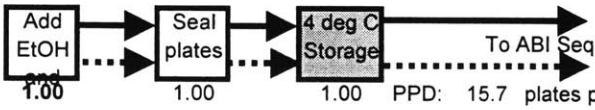

37 deg C warm
8.00
5
960
965
0.63
965
PPD: 15.7 plates per day
PLPD: 5.0 process labor hours per day
RT: 17.8 residence time, hours
This part of the spreadsheet uses batch sizes, set-up times, process times and associated labor to calculate total labor per step per day and theoretical process residence time.
DS= double-stranded DNA (pUC's)
The shaded areas are variables that can be changed to study effects.
-20 deg C Storage
PPD: 15.7 plates per day
PLPD: 1.7 process labor hrs per day
RT: 0.8 residence time, hours
8.00
5
5
10
1.25
10
Take pictures
Doc DNA/ Prep
-20 deg C Storage
To Sequencing Step
16.0
5
5
10
0.63
10
16.0
5
45
50
3.13
50
16.00
5
5
10
0.63
10
PPD: 15.7 plates per day
PLPD: 2.5 process labor hours per day
QLPD 2.7 QC labor hours per day
RT: 4.3 residence time, hours

Exhibit 2 (continued): Whitehead Institute Process Flow Diagram - Core Sequencing Operations



1.0	1.0	1.0	PPD:	36.7	plates per day
0.00	**0.00**	**0**	PLPD:	8.7	process labor hours per day
0.50	0.50	**5**	RT:	2.8	residence time, hours
0.50	0.50	5			
0.00	0.00	5.00			
0.50	0.50	5			



1.00	1.00	1.00	PPD:	15.7	plates per day
0.00	**0.00**	**5**	PLPD:	5.2	process labor hours per day
0.50	0.50	**5**	RT:	3.3	residence time, hours
0.50	0.50	10			
0.00	0.00	10.00			
0.50	0.50	10			

1.00	1.00	1.00	PPD:	10.5	plates per day
0.00	**0.00**	**5**	PLPD:	5.3	process labor hours per day
0.50	0.50	**5**	RT:	4.6	residence time, hours
0.50	0.50	10			
0.00	0.00	10.00			
0.50	0.50	10			

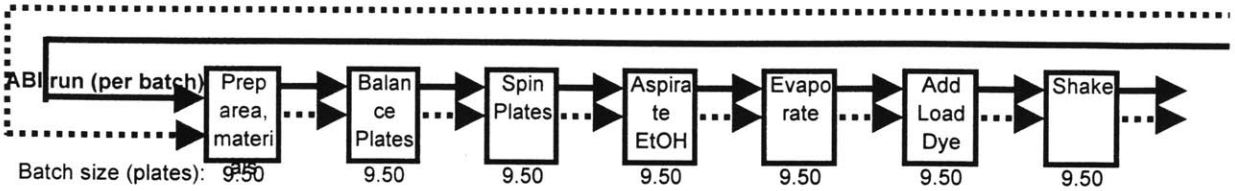
1.00	1.00	1.00	PPD:	0.0	plates per day
0.00	**0.00**	**5**	PLPD:	0.0	process labor hours per day
0.50	0.50	**5**	RT:	0.0	residence time, hours
0.50	0.50	10			
0.00	0.00	10.00			
0.50	0.50	10			

This part of the spreadsheet uses batch sizes, set-up times, process times and associated labor to calculate total labor per step per day and theoretical process residence time.

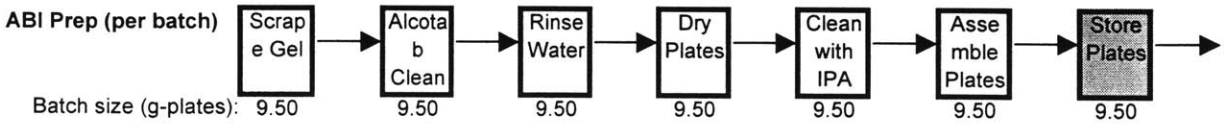
Only 3 of 4 ways of sequencing DNA are actually used.

The shaded areas are variables that can be changed to study effects.

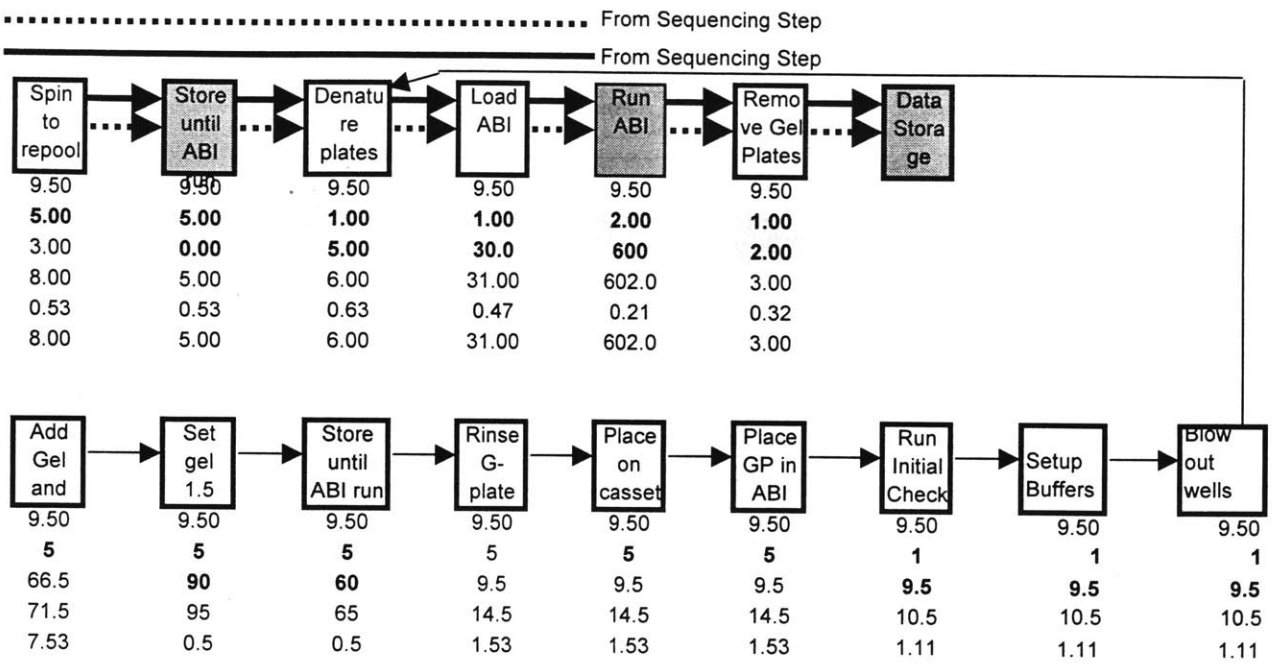
Exhibit 2 (continued): Whitehead Institute Process Flow Diagram - Core Sequencing Operations



	Prep area, materials	Balance Plates	Spin Plates	Aspirate EtOH	Evaporate	Add Load Dye	Shake
Batch size (plates):	9.50	9.50	9.50	9.50	9.50	9.50	9.50
Set-up time/batch:	**5**	**5.00**	**5.00**	**5.00**	**5.00**	**5.00**	**5.00**
Process time/batch:	**30**	9.50	**30.0**	9.50	**30.0**	9.50	9.50
Cycle time/batch:	35	14.50	35.00	14.50	35.00	14.50	14.50
Labor time/plate:	3.68	1.53	0.53	1.53	0.53	1.53	1.53
Residence time/plate:	0	14.50	35.00	14.50	35.00	14.50	14.50



	Scrape Gel	Alcotab Clean	Rinse Water	Dry Plates	Clean with IPA	Assemble Plates	Store Plates
Batch size (g-plates):	9.50	9.50	9.50	9.50	9.50	9.50	9.50
Set-up time/batch:	**5**	**10**	**5**	**5**	**5**	**5**	**5**
Process time/batch:	19	**30**	9.5	**30**	19	19	2.375
Cycle time/batch:	24	40	14.5	35	24	24	7.375
Labor time/g-plate:	2.53	1.05	1.53	0.53	2.53	2.53	0.78



This part of the spreadsheet uses batch sizes, set-up times, process times and associated labor to calculate total labor per step per day and theoretical process residence time.

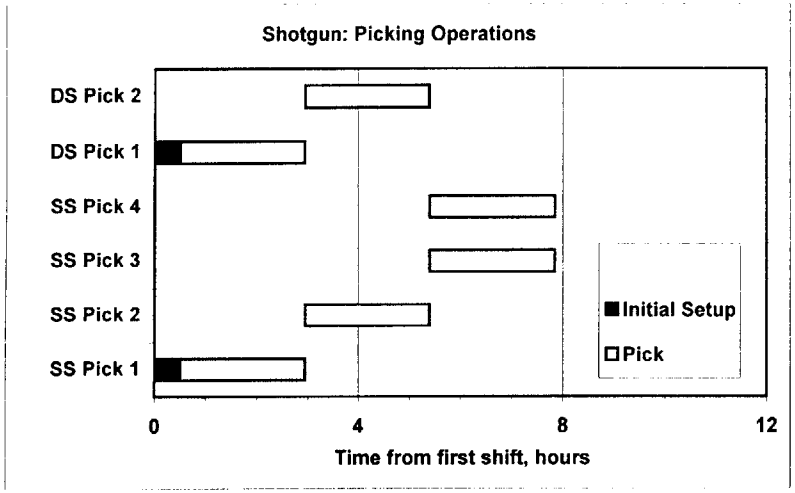
ABI run is the sample prep and loading of machines. ABI prep involves preparing the machines to run

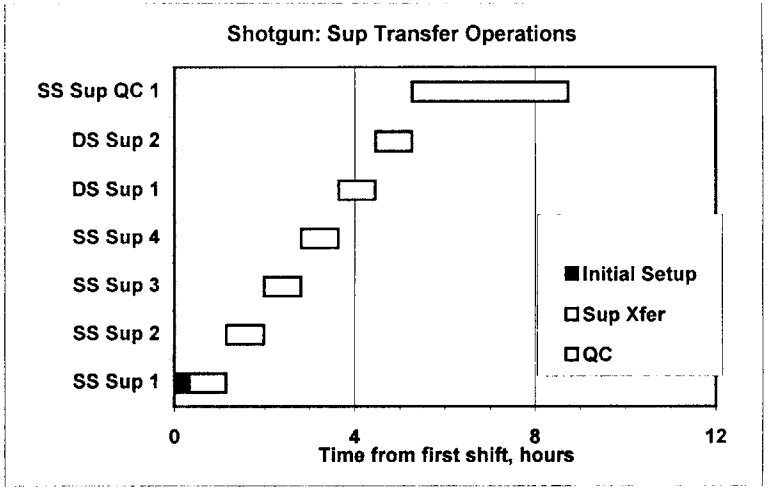
The shaded areas are variables that can be changed to study effects.

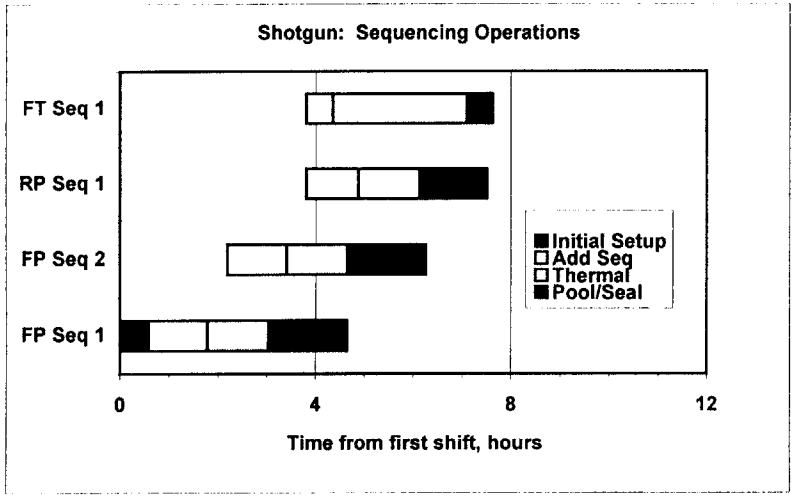
PPD:	76.0	plates per day
ALPD:	27.1	ABI run labor hours per day
PLPD:	25.4	ABI prep labor hours per day
RT:	13.1	residence time, hours

Exhibit 3: Whitehead Institute Scheduling Summary - Core Sequencing Operations (June, 1998)

Scenario: **63 SG plates/day** **40 ABI machines** **29 plates/project**







▲ **This part of the spreadsheet plots out a schedule based on the data from the following page.**
As can be seen, the three steps can be accomplished in one shift (8 hours).

Exhibit 3 (continued): Whitehead Institute Scheduling Summary - Core Sequencing Operations (June, 1998)

Picking

SS batches/shift:	3.931	Batch size=	8.00	plates
DS batches/shift:	1.966	Batch size=	8.00	plates

	Initial Setup	Pick
SS Pick 1	0.50	2.45
SS Pick 2	2.95	2.45
SS Pick 3	5.40	2.45
SS Pick 4	5.40	2.45
DS Pick 1	0.50	2.45
DS Pick 2	2.95	2.45

Supernatant Transfer

SS batches/	3.931	Batch size=	8.00	plates
DS batches/	1.966	Batch size=	8.00	plates
QC batches	1.05	Batch size:	30	plates

	Initial Setup	Sup Xfer	QC
SS Sup 1	0	0.33	0.83	
SS Sup 2	1.16		0.83	
SS Sup 3	1.99		0.83	
SS Sup 4	2.82		0.83	
DS Sup 1	3.64		0.81	
DS Sup 2	4.46		0.81	
SS Sup QC	5.27			3.48

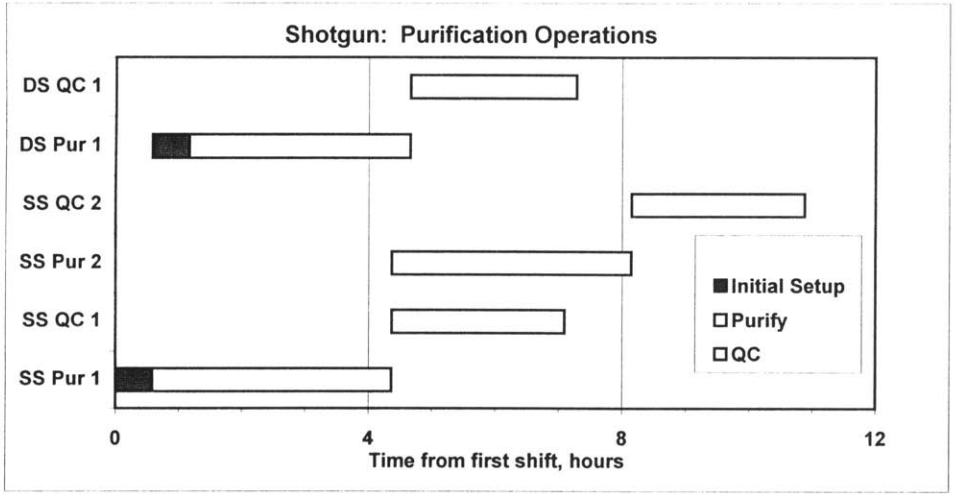
Primer Sequencing Setups

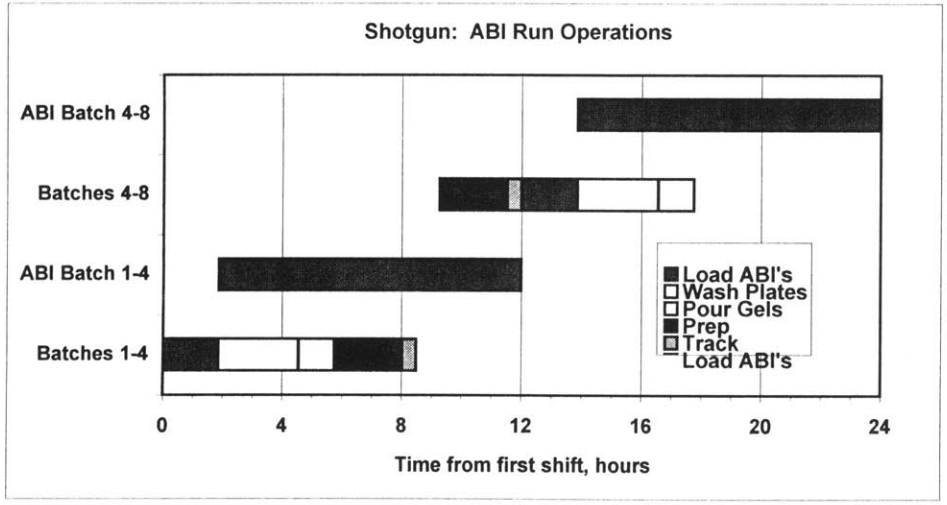
F prim batches/day:	1.93	Batch size=	19	plates		
R prim batches/day	0.98	Batch size=	16	plates		
F term batches/day	1.31	Batch size=	8	plates=	2	TC plates

	Initial Setup	Add Seq	Thermal	Pool/Seal
FP Seq 1	0	0.58	1.21	1.25	1.61
FP Seq 2	2.19		1.21	1.25	1.61
RP Seq 1	3.80		1.08	1.25	1.38
FT Seq 1	3.80		0.55	2.75	0.53

▲ **This part of the spreadsheet calculates a schedule based on data from the process flow diagram. This data is used to generate the plots on the previous page.**

Exhibit 3 (continued): Whitehead Institute Scheduling Summary - Core Sequencing Operations (June, 1998)





▲ **This part of the spreadsheet plots out a schedule based on the data from the following page.**
Purification can be accomplished in 8 hous, while the ABI sequencing machines run 24 hours per day.

Exhibit 3 (continued): Whitehead Institute Scheduling Summary - Core Sequencing Operations (June, 1998)

Purification

SS batches/s	2.097	Batch size=	15.00	plates
DS batches/	0.983	Batch size=	16.00	plates

		Initial Setup	Purify	QC
SS Pur 1	0	0.58	3.78	
SS QC 1	4.37			2.73
SS Pur 2	4.37		3.78	
SS QC 2	8.15			2.73
DS Pur 1	0.58	0.58	3.50	
DS QC 1	4.67			2.62

▲ **This part of the spreadsheet calculates a schedule based on data from the process flow diagram. This data is used to generate the plots on the previous page.**

ABI Run

Batches/shif	4.0	Batch size=	9.50	plates

		Load ABI's	Wash Plates	Pour Gels	Prep	Track	Load ABI's	Wash Plates	Pour Gels
Batches 1-4	0	1.87	2.69	1.19	2.27	0.48			
ABI Batch 1	1.87	10.13							
Batches 4-8	9.25	0.00	0.00	0.00	2.27	0.48	1.87	2.69	1.19
ABI Batch 4	13.87	10.13							

Exhibit 4: Whitehead Institute Statistical Measurements- Core Sequencing Operations

This is an example of the project tracking sheet used to coordinate operations.

Project	Size	Status	Comments	HO Entry Date	LC entry date	Pick Date	Prep Date	M13 Seq Date	M13	RR	pUCs	ABI	ABI Date
L252	200.00	GB				10/21/97			69/69	25/25	0/8	Roel	01/06/98
L258	105.00	GB				11/05/97			28/28	13/13	0/0		12/22/97
L285	95.00	GB				11/08/97			27/27	4/4	0/0		12/16/97
L221	145.00	SG				11/10/97			40/40	6/6	0/0		12/30/97
L279	50.00	Fin				11/11/97			44/44	3/3	0/6	Roel	01/07/98
L223	180.00	GB				11/14/97			44/44	n/a	0/0		12/23/97
L292	130.00	GB				11/14/97			35/35	4/4	0/0		12/11/97
L123	105.00	Assbly				11/15/97			54/54	n/a	0/0		12/15/97
L261	125.00	Fin				11/16/97			33/33	5/5	0/0		12/13/97
L111	130.00	Kill				11/19/97			46/46	7/7	0/0		12/24/97
L113	140.00	Kill				11/19/97			26/26	8/8	0/0		12/28/97
L294	190.00	Fin				11/24/97			54/54	1/1	0/0		12/12/97
L295	135.00	GB				11/25/97			41/41	n/a	0/0		12/19/97
L296	130.00	GB				12/02/97			40/40	n/a	0/0		12/30/97
L287	130.00	GB				12/05/97			39/39	6/6	0/0		12/23/97
L290	135.00	GB				12/06/97			42/42	4/4	0/0		01/06/98
L250	170.00	GB				12/09/97	12/23/97	12/30/97	46/46	5/5	0/0		01/09/98
L259	140.00	GB				12/16/97	12/30/97	1/6/98	34/34	7/7	0/0		01/15/98
L206	125.00	Dead				12/29/97	12/31/97	1/11/98	31/31	7/7	0/0		01/19/98
L257	97.00	GB				12/29/97	12/29/97	12/30/97	23/23	6/6	0/0		01/13/98
L300	110.00	GB				12/30/97	1/12/98	1/14/98	27/27	6/6	0/0		01/21/98
L302	97.00	GB				01/02/98	1/6/98	1/14/98	23/23	6/6	0/0		01/22/98
L306	125.00	Fin				01/06/98	1/8/98	1/15/98	31/31	7/7	0/0		02/25/98
L309	120.00	Fin				01/08/98	1/12/98	1/19/98	29/29	7/7	0/0		02/25/98
L195	40.00	Kill				01/08/98	1/15/98	1/19/98	12/12	2/2	0/0		01/28/98

Size of project, kilobases

KEY:
GB= Project finished, sent to GENBANK
SG= Project in Shotgun (Core Sequencing)
Fin= Project in Finishing Step
Kill=Project cancelled
L:QC=Project in Sequencing QC
L:LC=Project in Library Construction
HO=Project in Hand-Off (before Lib Const)

Prep is another word for the Purification Step

Exhibit 4 (cont'd): Whitehead Institute Statistical Measurements- Core Sequencing Operations, project data after assembly step

Project	Date of Assembly	Overall Pass Rate	Seq Pass Rate	Libr Pass Rate	Avg. Quality	Avg. Read Length	Length in Gap	% in Gap	Gap Data: Initial Assembly:Alewife	Gap Data: Initial Assembly: Phrap
L252	2/11/98	65	72	97	79	512	513	61	17	
L258	12/28/97	67	77	90	81	523	539	57	10	
L285	12/29/97	77	81	96	82	534	542	71	9	
L221		69	82	87	82	501				
L279	2/11/98	78	84	94	84	550	533	64	14	
L223	12/2297	71	81	90	83	541	555	41	16	
L292	12/22/97	83	89	95	87	560	562	75	11	
L123		65	75	91	79	499	436	25	13	
L261	12/22/97	73	83	89	83	529	536	72	12	
L111	8/20/96	51	64	87	70	489	472	28		
L113	9/96	53	59	94	66	456	371	28		
L294	12/18/97	74	79	96	82	533	534	75	12	
L295	12/25/97	85	89	96	87	562	563	77	5	
L296	12/31/97	86	88	98	86	555	569	84	1	
L287	12/26/97	86	89	97	86	527	546	86	9	
L290	01/09/98	80	82	93	84	560	559	69	7	
L250	01/15/98	76	83	93	83	554	531	77	10	
L259	03/10/98	79	82	97	84	576	547	75	13	
L206	03/10/98	67	76	92	80	534	419	7	28	
L257	01/21/98	79	82	97	83	560	549	77	4	
L300	01/29/98	76	82	94	83	546	534	74	4	
L302	02/15/98	77	83	94	83	552	532	77	4	
L306	03/15/98	76	81	95	83	577	544	70	11	
L309	03/13/98	76	81	95	83	574	545	73	8	
L195	03/05/98	82	88	93	85	554	524	71	9	

The Center eventually went to a new software for data assembly called "Phrap".

Exhibit 4: Whitehead Institute Statistical Measurements- Core Sequencing Operations
This is an example of the project tracking sheet used to coordinate operations.

Project	Size	Status	Comments	HO Entry Date	LC entry date	Pick Date	Prep Date	M13 Seq Date	M13	RR	pUCs	ABI	ABI Date
L311	135.00	GB				01/09/98	1/14/98	1/20/98	33/33	8/8	0/0		01/30/98
L228	130.00	GB				01/14/98	1/16/98	1/22/98	32/32	7/7	0/0		02/02/98
L310	105.00	GB				01/14/98	1/17/98	1/25/98	27/26	6/6	0/0		02/03/98
L291	137.00	Fin				01/16/98	1/20/98	1/29/98	33/33	8/8	0/0		02/04/98
L284	130.00	Fin				01/16/98	1/21/98	1/30/98	32/31	6/6	0/0		02/06/98
Two runs													
L293	110.00	Fin				01/22/98	1/23/98	2/2/98	27/27	4/4	2/2	33/33	02/09/98
L288	140.00	Fin				01/24/98	1/27/98	2/3/98	34/34	6/6	2/2	42/42	02/11/98
L297	155.00	GB				01/23/98	1/28/98	2/5/98	38/38	10/10	0/0	48/48	02/13/98
L283	125.00	GB				01/28/98	1/29/98	2/3/98	31/31	4/4	2/2	37/37	02/18/98
L215	155.00	GB				01/29/98	2/3/98	2/11/98	38/38	6/6	0/0	44/44	02/16/98
L286	140.00	Fin				01/30/98	2/5/98	2/16/98	34/34	6/6	2/2	42/42	02/22/98
L204	41.00	GB				01/28/98	2/5/98	2/17/98	10/10	2/2	0/0		02/23/98
L299	125.00	Fin				01/30/98	2/6/98	2/19/98	30/30	4/4	2/2	36/36	02/25/98
L301	140.00	Assbly	eed 4 plates F/R by ETs			02/01/98	2/9/98	2/23/98	34/34	6/6	6/6	42/42	02/27/98
L313	140.00	GB				02/02/98	2/11/98	2/25/98	34/34	6/6	2/2	42/42	04/06/98
L314	140.00	GB				02/04/98	2/12/98	3/2/98	34/34	6/6	2/2	42/42	04/06/98
L315	130.00	GB				02/05/98	2/18/98	3/5/98	34/34	6/6	2/2	42/42	04/07/98
L316	95.00	GB				02/09/98	2/20/98	3/9/98	23/23	4/4	2/2	27/27	04/08/98
L317	130.00	GB				02/09/98	2/23/98	3/10/98	31/31	6/6	0/0	37/37	04/09/98
L218	100.00	Fin				02/10/98	2/24/98	3/11/98	24/24	4/4	0/0	28/28	04/09/98
L312	110.00	Fin				02/12/98	2/26/98	3/12/98	26/26	4/4	0/0	30/30	04/10/98
L298	130.00	Fin				02/17/98	3/3/98	3/19/98	31/31	6/6	5/5	42/43	04/10/98
L303	100.00	Fin				02/13/98	3/10/98	3/22/98	24/24	4/4	4/4	32/32	03/26/98
L318	125.00	Fin				02/18/98	3/11/98	3/23/98	30/30	4/4	4/4	38/38	03/31/98
L319	160.00	Fin				02/23/98	3/18/98	3/25/98	40/40	6/6	6/6	52/52	04/02/98

Exhibit 4 (cont'd): Whitehead Institute Statistical Measurements- Core Sequencing Operations, project data after assembly step

Project	Date of Assembly	Overall Pass Rate	Seq Pass Rate	Libr Pass Rate	Avg. Quality	Avg. Read Length	Length in Gap	% in Gap	Gap Data: Initial Assembly:Alewife	Gap Data: Initial Assembly: Phrap
L311	02/10/98	74	84	90	84	602	589	70	6	
L228	02/18/98	66	75	91	80	571	556	66	11	
L310	02/15/98	75	79	96	81	569	553	73	4	
L291	02/18/98	70	77	94	80	596	581	63	12	
L284	02/19/98	71	79	92	82	613	578	69	5	
	Two runs									
L293	02/24/98	77	84	92	84	649	615	66	6	
L288	03/10/98	73	82	91	83	651	625	77	12	
L297	02/25/98	77	80	96	82	674	642	70	2	
L283	02/25/98	66	71	95	79	660	613	65	4	
L215	03/29/98	60	71	89	78	628	540	67	8	
L286	04/04/98	66	72	93	79	660	619	70	5	
L204	04/04/98	49	72	77	78	519	500	54	13	
L299	05/04/98	76	82	94	83	684	654	78	15	4
L301		68	78	90	81	637	604	65	24	
L313	04/17/98	74	79	95	82	676	639	66	4	
L314	04/18/98	76	80	96	82	665	630	74	5	
L315	04/16/98	78	83	95	84	691	650	70	3	
L316	04/23/98	78	80	97	84	699			0	
L317	04/18/98	78	80	97	83	702	655	73	6	10
L218	04/18/98	72	77	95	81	645	609	72	7	1
L312	04/16/98	72	77	96	81	686	644	65	4	3
L298	05/05/98	79	85	94	85	691	651	80	12	9
L303	05/25/98	67	73	94	77	611			13	5
L318	05/14/98	71	80	92	81	664	645	55	10	?
L319	05/14/98	79	83	95	81	655	663	63	9	2

Exhibit 4: Whitehead Institute Statistical Measurements- Core Sequencing Operations
This is an example of the project tracking sheet used to coordinate operations.

Project	Size	Status	Comments	HO Entry Date	LC entry date	Pick Date	Prep Date	M13 Seq Date	M13	RR	pUCs	ABI	ABI Date
New Coverage Rule: including vector 20 F 4R per 100 kb													
L172	40.00	GB				02/24/98	3/23/98	3/25/98	8/8	2/2	0/0	10/10	04/02/98
L308	120.00	Fin				02/25/98	3/24/98	3/29/98	25/25	4/4	4/4	33/33	04/03/98
L213	160.00	Fin				03/12/98	3/26/98	4/5/98	33/33	6/6	0/6	39/45	04/09/98
L324	123.00	GB	pUCs not in asmb stats			03/16/98	4/1/98	4/13/98	24/26	4/4	4/4	32/34	04/21/98
L289	125.00	Fin				03/19/98	3/27/98	4/15/98	26/28	0/0	4/4	30/32	04/23/98
L321	100.00	Assbly				03/23/98	3/30/98	4/17/98	21/21	0/0	4/4	25/25	04/28/98
L326	115.00	Fin	pUCs not in asmb stats			03/26/98	4/7/98	4/21/98	24/24	0/0	4/4	28/28	04/29/98
L327	103.00	Fin				03/27/98	4/8/98	4/21/98	19/20	0/0	4/4	23/24	04/30/98
L322	120.00	Fin				03/30/98	4/9/98	4/16/98	25/25	0/0	4/4	29/29	04/17/98
L330	105.00	Fin				04/07/98	4/21/98	5/3/98	22/21	0/0	4/4	26/25	05/05/98
L334	158.00	SG	ABI 6/15			04/09/98	4/22/98	4/30/98	32/32	0/0	6/6	38/38	05/04/98
L333	135.00	Fin				04/10/98	4/29/98	5/1/98	28/30	0/0	1/6	34/36	05/06/98
L323	85.00	SG2	ABI 6/12			04/13/98	4/30/98	5/4/98	20/17	0/0	4/4	24/21	05/07/98
L329	95.00	SG	ABI 6/12			04/15/98	5/5/98	5/11/98	25/20	0/0	4/4	29/24	05/19/98
change to DH5a													
L325	105.00	SG2	ABI 6/12		04/08/98	05/06/98	5/7/98	5/10/98	25/24	0/0	4/4	29/28	05/12/98
L332	115.00	SG			04/08/98	05/07/98	5/8/98	5/10/98	25/25	0/0	10/10	24/29	05/13/98
L335	119.00	SG			04/08/98	05/08/98	5/11/98	5/13/98	27/27	0/0	10/10	26/31	05/15/98
L337	137.00	Assbly			04/08/98	05/12/98	5/12/98	5/13/98	29/29	0/0	10/10	29/35	05/19/98
L343	170.00	SG			04/21/98	05/19/98	5/21/98	5/21/98	37/36	0/0	6/14	37/42	05/22/98
L255	140.00	SG			05/03/98	05/19/98	5/21/98	5/21/98	29/29	0/0	0/6	29/31	05/27/98
L338	180.00	SG			04/21/98	05/21/98	5/22/98	5/26/98	38/38	0/0	6/6	30/44	05/29/98
L339	200.00	SG			04/21/98	05/21/98	5/26/98	6/1/98	42/42	0/0	0/8	6/50	06/03/98
L341	160.00	SG			04/21/98	05/26/98	5/29/98	6/3/98	22/34	0/0	0/6	0/40	06/04/98
L340	190.00	SG			05/03/98	05/28/98	6/1/98	6/3/98	24/40	0/0	0/6	0/46	06/05/98
L328	80.00	SG			05/05/98	05/29/98	6/2/98	6/4/98	12/19	0/0	0/4	0/23	06/05/98
L346	190.00	SG		05/01/98	05/07/98	06/01/98	6/4/98	6/4/98	22/40	0/0	0/8		06/08/98

Exhibit 4 (cont'd): Whitehead Institute Statistical Measurements- Core Sequencing Operations, project data after assembly step

Project	Date of Assembly	Overall Pass Rate	Seq Pass Rate	Libr Pass Rate	Avg. Quality	Avg. Read Length	Length in Gap	% in Gap	Gap Data: Initial Assembly:Alewife	Gap Data: Initial Assembly: Phrap
New Coverage Rule: including vector 20 F 4R per 100 kb										
L172	04/17/98	77	83	94	84	687	632	77	0	
L308	05/08/98	69	73	96	79	658	555	64	7	5
L213	05/01/98	72	83	89	84	699	665	68	15	3
L324	05/19/98	74	78	96	80	659	652	56	8	3
L289	05/15/98	70	78	92	79	662			7	1
L321	06/12/98	66	71	95	76	637				
L326	05/20/98	72	79	93	75	639	660	51	10	?
L327	05/28/98	69	74	95	63	699			9	**1**
L322	05/27/98	74	81	92	75	715			7	**2**
L330	06/07/98	77	87	90	71	710			**4**	?
L334		69	75	94	71	724				
L333	06/08/98	76	84	93	72	714			11	**4**
L323	06/01/98	72	84	89	66	716			**3**	1
L329		75	85	90	66	765				
change to DH5a										
L325		78	86	92	59	722				
L332		78	87	92	54	779				
L335		89	95	94	51	805				
L337	06/12/98	89	95	94	51	801				
L343		89	94	95	43	798				
L255		88	93	95	48	740				
L338		88	96	92	45	784				
L339		96	98	98	67	783				
L341										
L340										
L328										
L346										

Exhibit 4: Whitehead Institute Statistical Measurements- Core Sequencing Operations
This is an example of the project tracking sheet used to coordinate operations.

Project	Size	Status	Comments	HO Entry Date	LC entry date	Pick Date	Prep Date	M13 Seq Date	M13	RR	pUCs	ABI	ABI Date
New Coverage 12 M13 8Pr, 4T; 12 pUC, 6F:6R Pr per 100kb, incl vector													
L347	175.00	SG		05/01/98	05/07/98	06/03/98	6/5/98	6/8/98	15/22		0/11		06/10/98
L345	165.00	SG		05/01/98	05/14/98	06/08/98	6/9/98	6/9/98	13/21		0/10		
L336	190.00	SG		03/19/98	5/19/98	06/10/98	6/10/98	6/12/98	16/24		0/12		
L348	180.00	SG		05/01/98		06/08/98	6/9/98	6/11/98	16/23		0/11		
L350	200.00	SG		05/11/98	5/21/98	06/11/98			0/25		0/13		
L351	185.00	SG		05/11/98	5/21/98				0/23		0/11		
L352	170.00	SG		05/11/98	5/21/98				0/21		0/10		
L354	165.00	L:QC	pUCs only, McPrep!!	05/11/98	5/26/98				0/21		0/10		
L356	180.00	L:QC	electroporation	05/11/98	5/26/98				0/23		0/11		
L342	200.00	L:LC		04/09/98	6/8/98								
L344	190.00	L:LC		05/14/98	6/8/98								

Exhibit 4 (cont'd): Whitehead Institute Statistical Measurements- Core Sequencing Operations, project data after assembly step

Project	Date of Assembly	Overall Pass Rate	Seq Pass Rate	Libr Pass Rate	Avg. Quality	Avg. Read Length	Length in Gap	% in Gap	Gap Data: Initial Assembly:Alewife	Gap Data: Initial Assembly: Phrap
New Coverage 12 M13 8Pr, 4T; 12 pUC, 6F:6R Pr per 100kb, incl vector										
L347										
L345										
L336										
L348										
L350										
L351										
L352										
L354										
L356										
L342										
L344										

Exhibit 5- Regression of B/W(Wc/Bc) vs. %SV

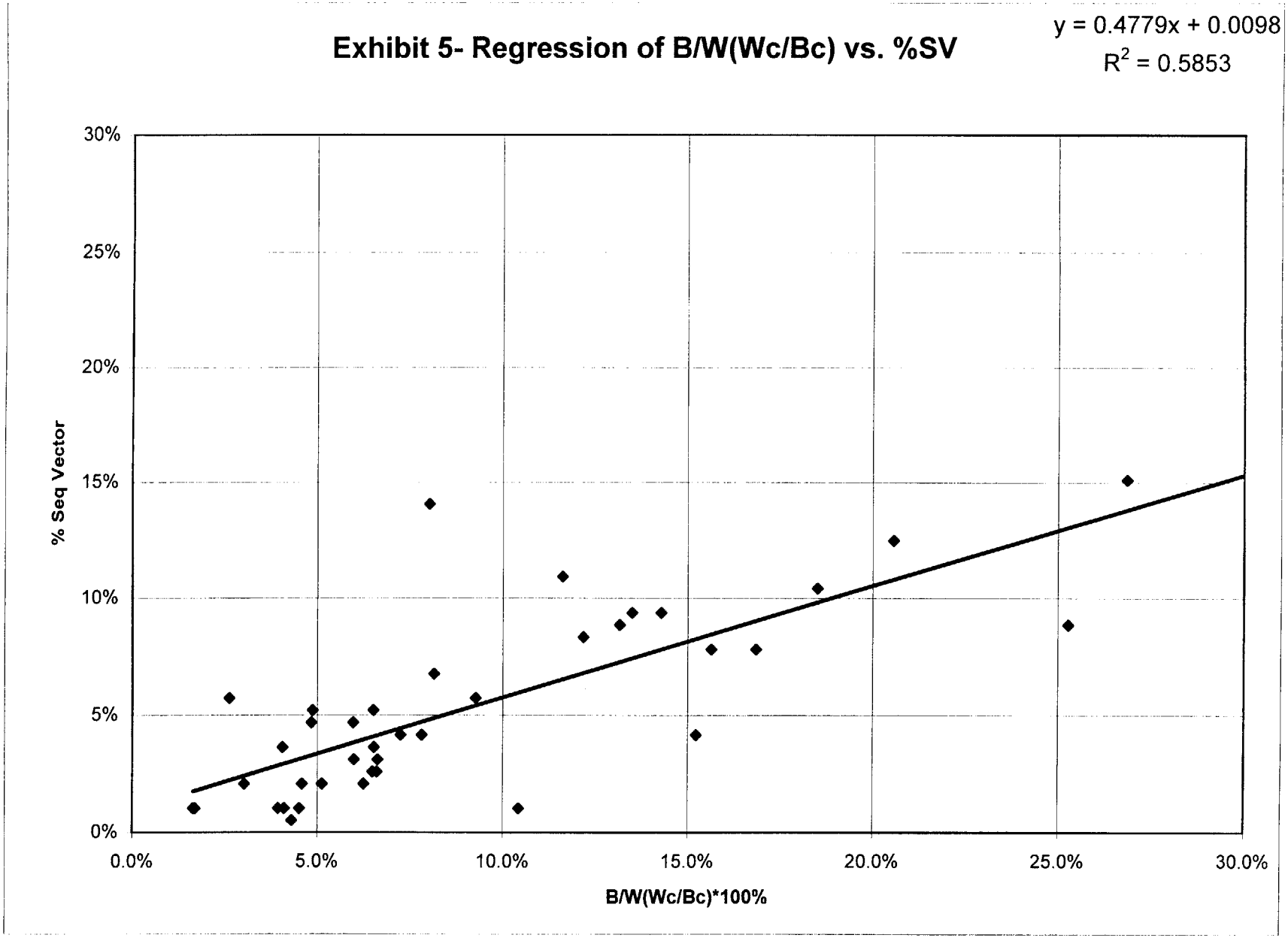


Exhibit 6- Regression of B/W(Wc/Bc) vs. %SV+SI

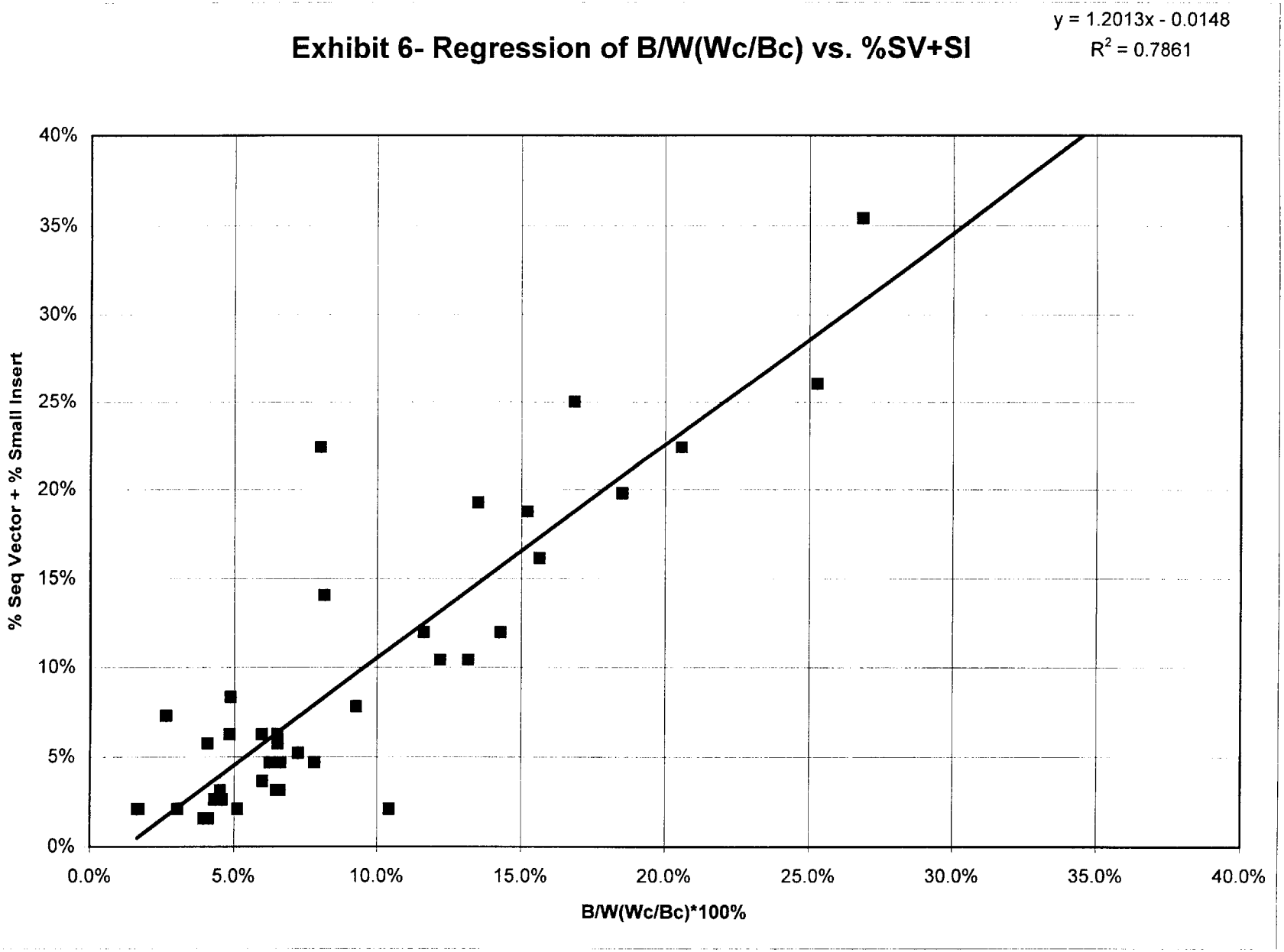


Exhibit 7: Operating Characteristic Curves

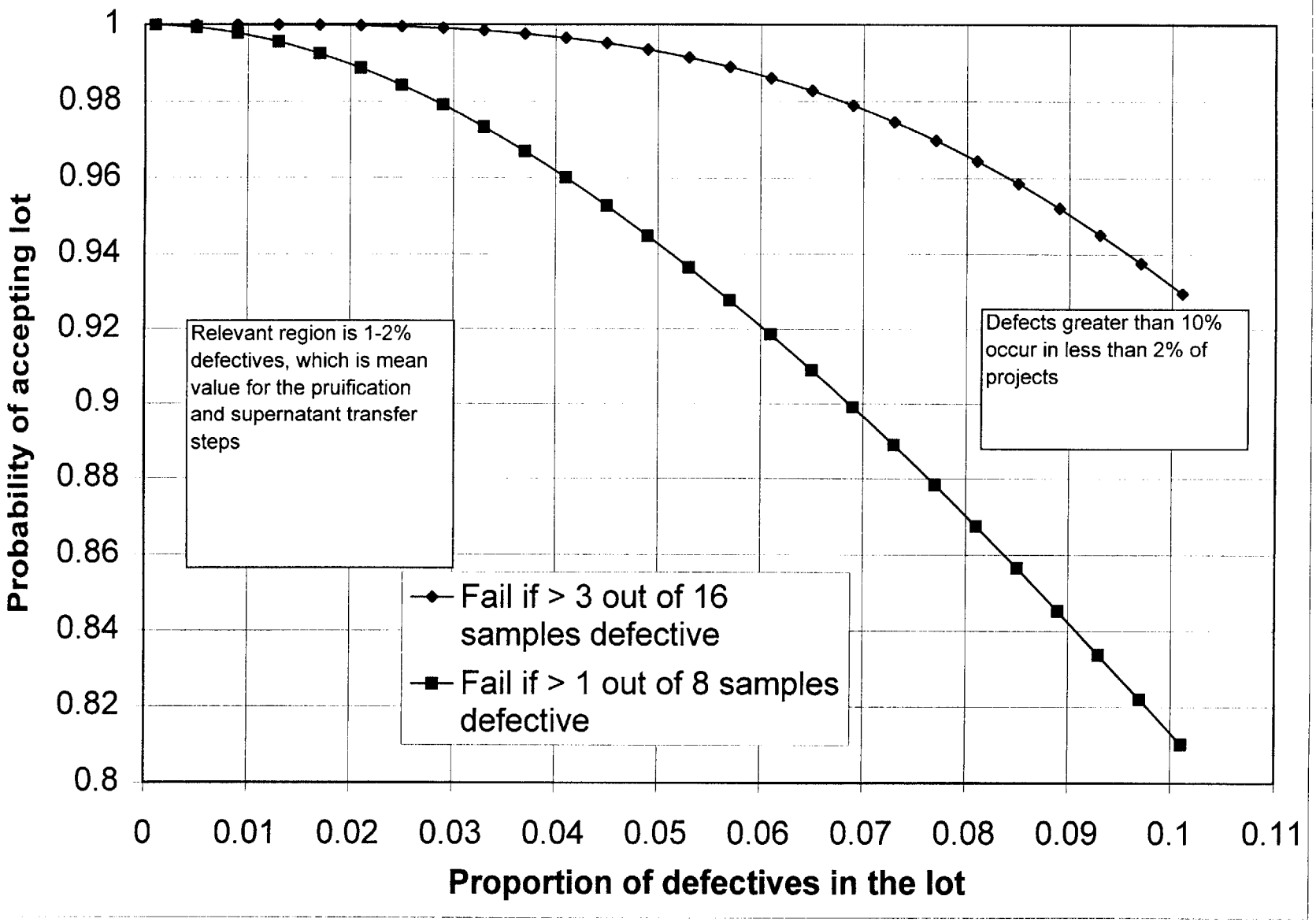


Exhibit 8 Whitehead Institute - Library Construction Sequencing QC

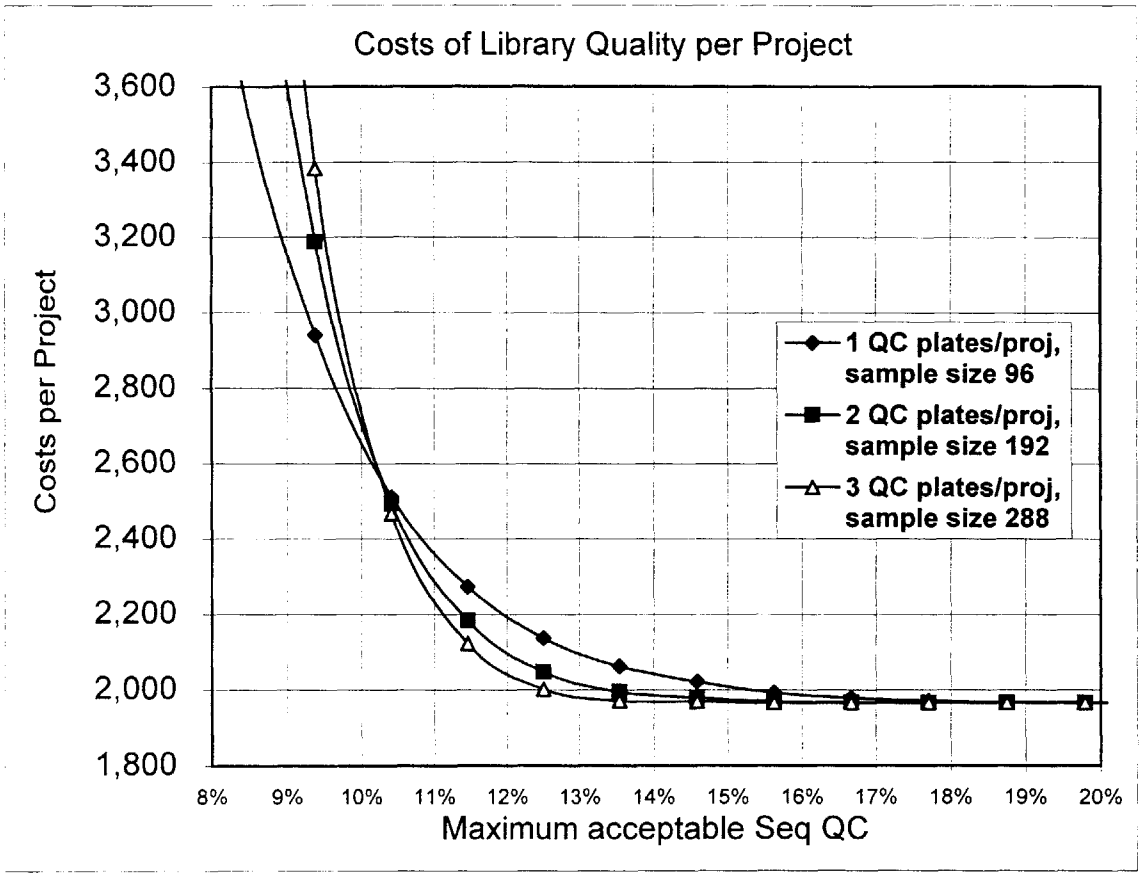
Seq QC Cost Model

Mean project quality w/o QC:	9.5%	
Plates/proj:	30	
New Library Cost:	$ 1,325	per library
Cost of Seq:	**$ 225**	per plate
% Lost QC Data:	**50%**	

Total No. Test Wells	Max No. Well Fails	% Lib Fails	Proj Qual	Fraction accepted		Library costs/proj: Quality	Generate	Lost Data	Total Lib Costs
96	4.0	4.2%	4.0%	0.04	$	270	30,524	2,479	**33,273**
96	5.0	5.2%	5.1%	0.10	$	346	13,647	1,046	**15,039**
96	6.0	6.3%	5.8%	0.18	$	389	7,258	504	**8,151**
96	7.0	7.3%	6.5%	0.30	$	436	4,448	265	**5,149**
96	8.0	8.3%	7.0%	0.43	$	475	3,063	148	**3,686**
96	9.0	9.4%	7.9%	0.57	$	536	2,321	85	**2,941**
96	10.0	10.4%	8.3%	0.70	$	562	1,901	49	**2,512**
96	11.0	11.5%	8.8%	0.80	$	591	1,655	28	**2,273**
96	12.0	12.5%	9.1%	0.88	$	612	1,509	16	**2,137**
96	13.0	13.5%	9.3%	0.93	$	630	1,425	8	**2,063**
96	14.0	14.6%	9.5%	0.96	$	641	1,376	4	**2,022**
96	15.0	15.6%	9.5%	0.98	$	641	1,350	2	**1,994**
96	16.0	16.7%	9.5%	0.99	$	641	1,337	1	**1,979**
96	17.0	17.7%	9.5%	1.00	$	641	1,330	0	**1,972**
96	18.0	18.8%	9.5%	1.00	$	641	1,327	0	**1,969**
96	19.0	19.8%	9.5%	1.00	$	641	1,326	0	**1,967**
96	20.0	20.8%	9.5%	1.00	$	641	1,325	0	**1,967**
96	21.0	21.9%	9.5%	1.00	$	641	1,325	0	**1,966**
96	22.0	22.9%	9.5%	1.00	$	641	1,325	0	**1,966**
96	23.0	24.0%	9.5%	1.00	$	641	1,325	0	**1,966**
96	24.0	25.0%	9.5%	1.00	$	641	1,325	0	**1,966**

Total No. Test Wells	Max No. Well Fails	% Lib Fails	Proj Qual	Fraction accepted		Library costs/proj: Quality	Generate	Lost Data	Total Lib Costs
192	8.0	4.2%	4.0%	0.00	$	270	295,905	50,023	**346,198**
192	10.0	5.2%	5.1%	0.02	$	346	60,733	10,088	**71,167**
192	12.0	6.3%	5.8%	0.07	$	389	18,156	2,858	**21,403**
192	14.0	7.3%	6.5%	0.18	$	436	7,372	1,027	**8,835**
192	16.0	8.3%	7.0%	0.34	$	475	3,847	428	**4,751**
192	18.0	9.4%	7.9%	0.54	$	536	2,460	193	**3,188**
192	20.0	10.4%	8.3%	0.72	$	562	1,844	88	**2,494**
192	22.0	11.5%	8.8%	0.85	$	591	1,554	39	**2,184**
192	24.0	12.5%	9.1%	0.93	$	612	1,419	16	**2,048**
192	26.0	13.5%	9.3%	0.97	$	630	1,360	6	**1,996**
192	28.0	14.6%	9.5%	0.99	$	641	1,336	2	**1,980**
192	30.0	15.6%	9.5%	1.00	$	641	1,328	1	**1,970**
192	32.0	16.7%	9.5%	1.00	$	641	1,326	0	**1,967**
192	34.0	17.7%	9.5%	1.00	$	641	1,325	0	**1,966**
192	36.0	18.8%	9.5%	1.00	$	641	1,325	0	**1,966**
192	38.0	19.8%	9.5%	1.00	$	641	1,325	0	**1,966**
192	40.0	20.8%	9.5%	1.00	$	641	1,325	0	**1,966**
192	42.0	21.9%	9.5%	1.00	$	641	1,325	0	**1,966**
192	44.0	22.9%	9.5%	1.00	$	641	1,325	0	**1,966**
192	46.0	24.0%	9.5%	1.00	$	641	1,325	0	**1,966**
192	48.0	25.0%	9.5%	1.00	$	641	1,325	0	**1,966**

Exhibit 8 (continued) Whitehead Institute - Library Construction Sequencing QC



Total No. Test Wells	Max No. Well Fails	% Lib Fails	Proj Qual	Fraction accepted	Library costs/proj: Quality	Generate	Lost Data	Total Lib Costs
288	12	4.2%	4.0%	0.00	270	2,554,481	650,332	**3,205,083**
288	15	5.2%	5.1%	0.01	346	242,947	61,545	**304,839**
288	18	6.3%	5.8%	0.03	389	41,374	10,201	**51,964**
288	21	7.3%	6.5%	0.12	436	11,334	2,549	**14,319**
288	24	8.3%	7.0%	0.29	475	4,584	830	**5,890**
288	27	9.4%	7.9%	0.52	536	2,538	309	**3,383**
288	30	10.4%	8.3%	0.74	562	1,787	118	**2,468**
288	33	11.5%	8.8%	0.89	591	1,490	42	**2,123**
288	36	12.5%	9.1%	0.96	612	1,376	13	**2,002**
288	39	13.5%	9.3%	0.99	630	1,338	3	**1,971**
288	42	14.6%	9.5%	1.00	641	1,328	1	**1,970**
288	45	15.6%	9.5%	1.00	641	1,325	0	**1,967**
288	48	16.7%	9.5%	1.00	641	1,325	0	**1,966**
288	51	17.7%	9.5%	1.00	641	1,325	0	**1,966**
288	54	18.8%	9.5%	1.00	641	1,325	0	**1,966**
288	57	19.8%	9.5%	1.00	641	1,325	0	**1,966**
288	60	20.8%	9.5%	1.00	641	1,325	0	**1,966**
288	63	21.9%	9.5%	1.00	641	1,325	0	**1,966**
288	66	22.9%	9.5%	1.00	641	1,325	0	**1,966**
288	69	24.0%	9.5%	1.00	641	1,325	0	**1,966**
288	72	25.0%	9.5%	1.00	641	1,325	(0)	**1,966**

Exhibit 9: Whitehead Institute- Core Sequencing Operations
Histogram of Picked Plates per Day
Sample Size = 110 days, Mean = 20

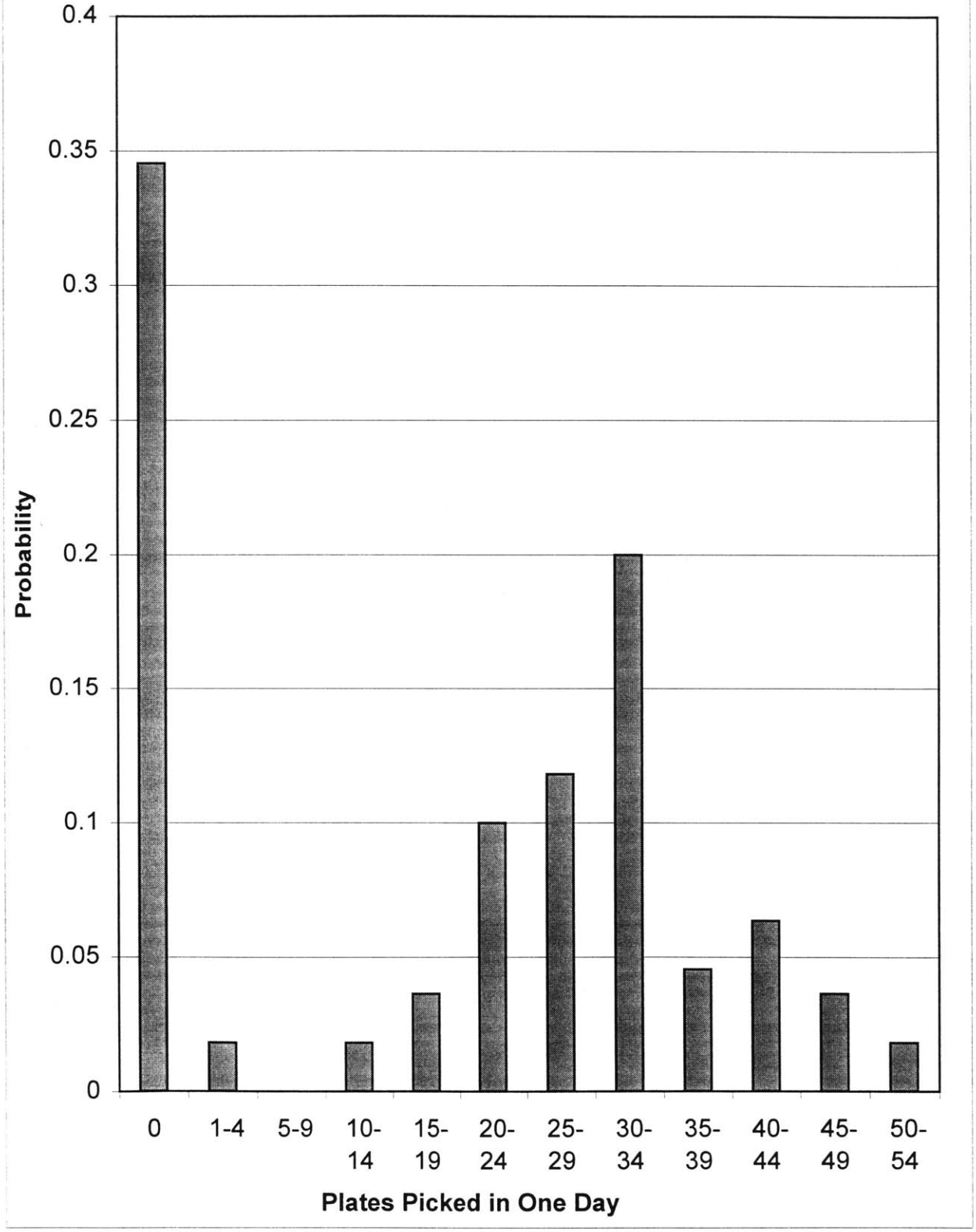


Exhibit 10: Whitehead Institute- Core Sequencing Operations
Histogram of Purified Plates per Day
Sample Size = 110 days, Mean = 18

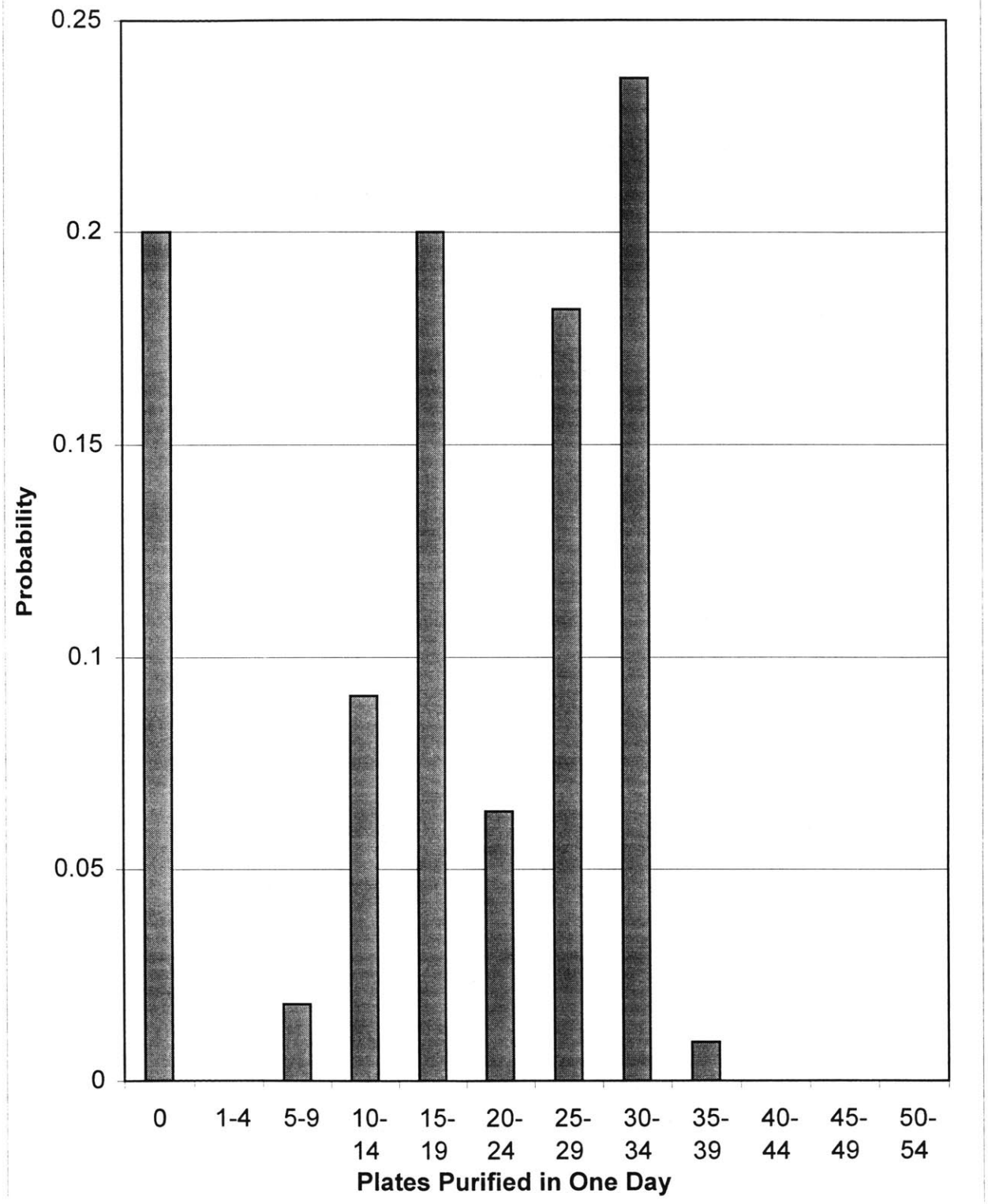


Exhibit 11: Whitehead Institute- Core Sequencing Operations
Histogram of Sequenced Plates per Day
Sample Size = 110 days, Mean = 21

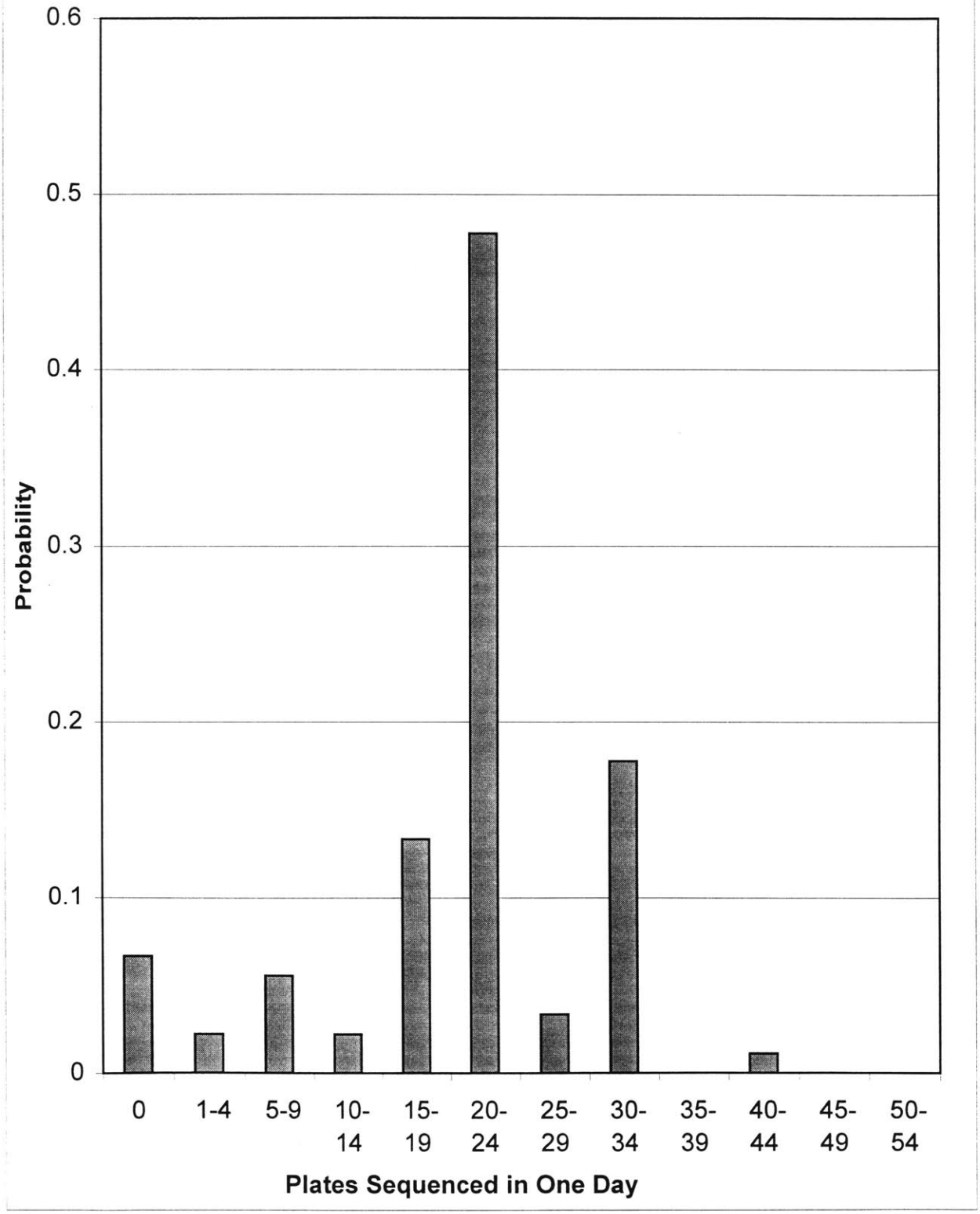


Exhibit 12: Whitehead Institute- Core Sequencing Operations
Histogram of Plates Run on ABI per Day
Sample Size = 25 days, Mean = 28

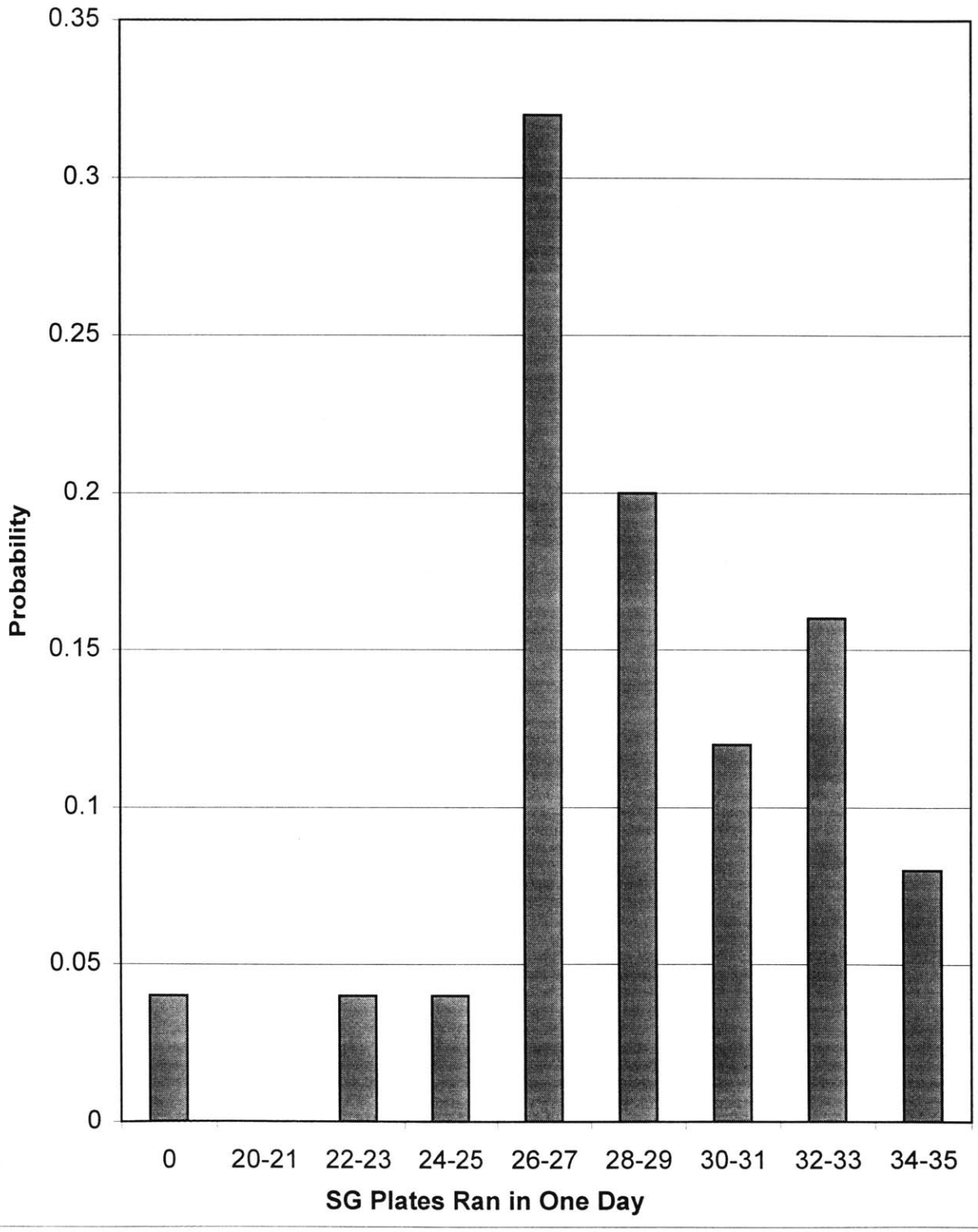


Exhibit 13: Whitehead Institute Stochastic Simulation of Inventories for Core Sequencing

Min Inv:	60	40	80		**Mean Throughput Rates:**				**Throughput "Multipliers"**			
Max Inv:	80	60	120		30	30	24	20				
Date	**Sup Inv**	**Pur Inv**	**Seq Inv**	**ABI Inv**	**Sup**	**Pur**	**Seq**	**ABI**	**Sup**	**Pur**	**Seq**	**ABI**
1	70	50	100	0	30	30	24	20	1	1	1	1
2	70	56	104	20	30	30	24	20	1	1	1	1
3	70	62	108	40	30	0	24	20	1	1	1	1
4	100	38	112	60	0	30	0	20	1	1	1	1
5	70	68	92	80	30	0	24	20	1	1	1	1
6	100	44	96	100	0	30	24	20	1	1	1	1
7	70	50	100	120	30	30	24	20	1	1	1	1
8	70	56	104	140	30	30	24	20	1	1	1	1
9	70	62	108	160	30	0	24	20	1	1	1	1
10	100	38	112	180	0	30	0	20	1	1	1	1
11	70	68	92	200	30	0	24	20	1	1	1	1
12	100	44	96	220	0	30	24	20	1	1	1	1
13	70	50	100	240	30	30	24	20	1	1	1	1
14	70	56	104	260	30	30	24	20	1	1	1	1
15	70	62	108	280	30	0	24	20	1	1	1	1
16	100	38	112	300	0	30	0	20	1	1	1	1
17	70	68	92	320	30	0	24	20	1	1	1	1
18	100	44	96	340	0	30	24	20	1	1	1	1
19	70	50	100	360	30	30	24	20	1	1	1	1
20	70	56	104	380	30	30	24	20	1	1	1	1
21	70	62	108	400	30	0	24	20	1	1	1	1